

ANOTACIJA I DETEKCIJA NEGACIJE U KOMENTARIMA PROIZVODA**ANNOTATION AND DETECTION OF NEGATION IN PRODUCT REVIEWS**Aleksandra Milijević, *Fakultet tehničkih nauka, Novi Sad***Oblast – RAČUNARSTVO I AUTOMATIKA**

Kratak sadržaj – Rad je baziran na kreiranju sistema za anotaciju i detekciju negacije u komentarima razne vrste proizvoda na engleskom jeziku.

Ključne reči: *Mašinsko učenje, Negativne reči, Opseg negacije*

Abstract – *This paper presents the system for the annotation and detection of negation in the reviews of various types of products in the English language.*

Keywords: *Machine learning, Negation cues, Scope of negation*

1. UVOD

Poslednjih godina pojam negacije u prirodnom jeziku predstavlja sve veći objekat interesovanja. Detekcija negacije je važan deo u analizi sentimenata (eng. Sentiment Analysis) u oblasti Obrade prirodnog jezika (NLP) i smatra se jednim od najtežih problema za rešavanje.

Oblasti u kojima je detekcija negacije veoma bitna su klinički zapisi, komentari na društvenim mrežama, kupovina preko interneta, recenzije filmova. U slučaju prve oblasti zdravlje pacijenata zavisi od pravilno kreiranog sistema, dok u ostalim oblastima, samim korisnicima je veoma korisna informacija koji prizvod da kupe ili koji film da gledaju.

Vrste negacije koje postoje su neafiksne negacije sa jednom reči, afiksne negacije i negacije koje su iskazane pomoću više reči. U ovom radu su obrađene samo prve dve vrste.

Sistem kreiran u radu sastoji se iz sledećih delova: polu-automatske anotacije negativnih reči u rečenicama, detekcije negativnih reči, detekcije opsega negacije. U svakom od poslednja dva dela vrši se izbor atributa i klasifikacija uz pomoć algoritama mašinskog učenja. Za klasifikaciju negativnih reči koristi se Support Vector Machine (SVM), a za opseg negacije Conditional Random Fields (CRF) algoritam. Rezultati SVM algoritma evaluirani su pomoću kros-validacije i F1 mere, a CRF pomoću F1 mere.

2. PREGLED STANJA U OBLASTI

Postoji više pristupa koji mogu da se koriste za detekciju negacije. Najranija istraživanja kreirala su određena pravila pomoću kojih određuju negaciju u rečenicama (eng. Rule-based negation detection). Podaci koji su se koristili bili su najvećim delom iz medicinskog domena kao što su klinički izveštaji pacijenata. Rad [1] je jedan od prvih u kome je korišćen prethodno naveden pristup. Autori rada su razvili jednostavan algoritam regularnih izraza pod nazivom NegEx koji je postigao tačnost od 84.5%.

2007. godine u radu [2] kreiran je algoritam ConText koji predstavlja proširenje NegEx-a. Osim negacije vrši i određivanje da li su klinička stanja navedena u kliničkim izveštajima hipotetička, istorijska, ili ih je iskusio neko drugi a ne pacijent. 2008. godine [3] je objavio BioScope korpus medicinskih tekstova koji služe kao osnova za anotaciju korpusa negativnim rečima i opsega negacije prilikom detekcije negacije.

2009. objavljen je cTAKES algoritam koji nije uspevao da detektuje kompletne obrasce negacije. Iz tog razloga [4] je koristio parser zavisnosti za detekciju negacije čime su poboljšane performanse cTAKES sistema.

Autori rada [5] su 2012. godine objavili anotirani skup podataka sa negativnim rečima i opsegom negacije. Skup podataka sadrži poznate priče Konana Dojla i služi kao osnova za identifikovanje negacije u tekstovima.

U radu [6] određivalo se koje reči se negiraju ključnom reči negacije. Vršeno je poređenje više metoda za određivanje opsega negacije. Donesen je zaključak da se u najvećoj meri opseg negacije prostire u rasponu od 6 reči sa desne strane, dok sa leve stane u proseku od 3 reči.

U radu [7] korišćen je pristup mašinskog učenja za detekciju negacije. Koristili su CRF model koji je obučen na osnovu atributa ekstrahovanih pomoću parsera zavisnosti na BioScope i Product Reviews korpusu. U radu [8] opisan je prvi sofisticirani sistem za detekciju obima negacije prilikom analize sentimenata Twitter tekstova. [9] je predložio lingvistički model zasnovan na SVM i CRF algoritmu za detekciju negacije. SVM se koristi za klasifikaciju reči negacije na osnovu maksimalne margine, dok je detekcija opsega negacije vršena kao problem sekvencijalnog labelovanja korišćenjem CRF algoritma.

3. TEORIJSKI POJMOVI I DEFINICIJE**3.1. Definicija i primeri negacije**

Negacija je konverzija afirmativne rečenice koja pretvara prostu potvrdnu rečenicu u negativnu. Identifikuje se

NAPOMENA:

Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, red.prof.

pomoću znaka negacije, koji se sastoji od jednog ili više tokena koji signaliziraju negaciju.

Oni mogu biti od jedne reči, više reči, ili afiksalni kao što je prefiks *im-* u reči *impossible* u primeru 3.1.5, ili sufixs *-less* u *endless*, a u nekim slučajevima reči koje povezuju dve ili više negativnih alternativa, kao što su *neither/nor*. Postoji opseg negacije, gde su reči samo unutar tog opsega pogođene negacijom, a ne cela rečenica. Postoji situacije u kojima reči negacije nemaju opseg (3.1.3), ili je opseg dvosmisle kao u primeru 3.1.1. U daljim primerima znak negacije je označen podebljanim slovima, a opseg uglastim zagradama.

- Jedna reč: *No, nor, never, not*

3.1.1 [Anna is] **not** [short and fat].

3.1.2 We have two dogs, **neither** [one can sit].

3.1.3 **No**, we did not go to the market.

- Više reči: *no longer, by no means, for nothing*

3.1.4 [There is] **no longer** [snow outside].

- Afiksi: *im-, un-, dis-, ir-, in-, -less*

3.1.5 [Boys are] **impossible** [to understand].

U ovom radu uzete su u obzir samo afiksalne i neafiksne negacije sa jednom reči.

3.2. SVM algoritam

SVM je nadgledani algoritam za učenje koji se koristi za probleme klasifikacije i regresije. Može da rešava linearne

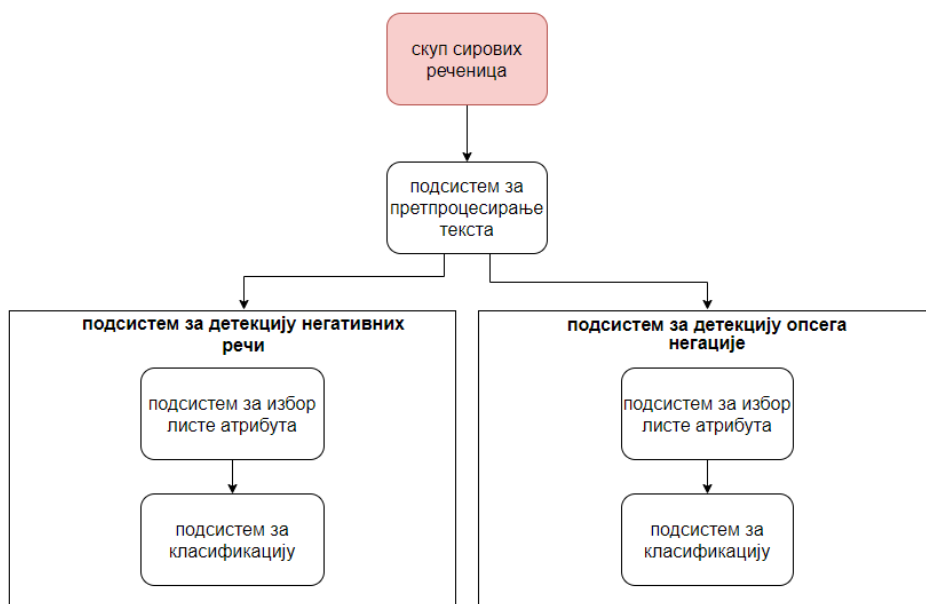
i nelinearne probleme i ideja je ta da algoritam kreira liniju ili hiperravan koja razdvaja podatke u klase. Bitan hiperparametar kod ovog algoritma je parametar regularizacije C i govori nam koliko pogrešne klasifikacije želimo da izbegnemo. Za velike vrednosti C javlja se tvrdi SVM i on ne dozvoljava greške u klasifikaciji (može izazvati overfitting). Dok za male vrednosti C javlja se meki SVM koji dozvoljava greške prilikom klasifikacije.

3.3. CRF algoritam

CRF je tipa diskriminativnog klasifikatora. Ulazni podaci su sekvencijalni i mora se uzeti u obzir prethodni kontekst kada se vrši predviđanje. Nedostatak je visoka računarska složenost faze obuke algoritma. CRF ne radi sa nepoznatim rečima, odnosno rečima koje nisu bile prisutne u uzorku podataka za obuku. Neke od prednosti: moguće je postići visok kvalitet označavanja ako se odaberu prave karakteristike; fleksibilan u pogledu izbora karakteristika itd. Jedni od bitnih hiperparametara su maksimalan broj iteracija, $c1$ i $c2$ koeficijenti za $L1$ i $L2$ regularizaciju respektivno.

4. METODOLOGIJA

Sistem se sastoji od tri podsistema: podsistema za preprocesiranje teksta, podsistema za detekciju negativnih reči i podsistema za detekciju opsega negacije, slika 1.



Slika 1. Arhitektura sistema

4.1. Podsystem za preprocesiranje

U prvom podsystemu vršeno je preprocesiranje rečenica uz pomoć zbirke alata Stanza gde su primenjene tehnike tokenizacije i lematizacije. Svaka rečenica je podeljena na reči i za svaku reč čuvane su sledeći atributi: token, lema, POS tag (univerzalni deo govora - part-of-speech), id glave, deprel (odnos zavisnosti između posmatrane reči i njene sintetičke glave). Takođe je vršena anotacija negativnih reči gde je sačuvana informacija da li je reč negativna ili ne. Postoje dva leksikona: leksikon negativnih reči i leksikon afiksa na osnovu kojih je određen poslednje naveden atribut (tabele 1. i 2.).

Tabela 1. Leksikon negativnih reči

hardly	neednt	isnt	without
neither	wasnt	wouldnt	no
nobody	lack	never	nowhere
not	nor	nothing	cant
cannot	none	dont	doesnt
didnt	n't	hasnt	havnt
havnt	hadnt	mightnt	shouldnt

Tabela 2. Leksikon negativnih prefiksa i sufiksa

un-	il-	anti-
in-	ir-	non-
im-	dis-	-less

4.2. Podsystemi za detekciju negativnih reči i opsega

Podsystemi za detekciju negativnih reči i opsega negacije dele se na dva dela: određivanje liste atributa, klasifikacija. Atributi izabrani kod prvog podsystema su: token, lema, POS tag, lema prethodne/naredne reči posmatranog tokena, POS tag prethodne/naredne reči posmatranog tokena, afiks (kod afiksne negacije), n-gram afikslnih negativnih reči (deo reči bez afiksa). Što se tiče podsystema za detekciju opsega negacije, atributi koji su korišćeni su isti kao i za prvi podsystem (bez atributa vezanih za afikse) sa dodatkom sintaksičkih atributa: veze zavisnosti (deprel), POS tag glave posmatranog tokena, usmerena putanja zavisnosti od glave negativne reči do svakog tokena (najkraći put od glave negativne reči do posmatranog tokena - korišćen Dijkstra algoritam). U fazi klasifikacije, SVM je korišćen za detekciju negativnih reči, dok za detekciju opsega je upotrebljen CRF. Oba algoritma su pokazala dobre rezultate u ranijim radovima koji su obrađivali ovu temu.

5. EKSPERIMENTI I REZULTATI

U radu su vršena dva eksperimenta. Prvi se odnosi na detekciju negativnih reči, i testirane su razne kombinacije atributa i vrednosti hiperparametara da bi se dobio optimalan rezultat. Drugi eksperiment je vezan za detekciju opsega negacije i takođe su vršene razne kombinacije atributa i vrednosti hiperparametara. Ulaz u oba podsystema je rečnik reči gde je za svaku reč označeno kojoj rečenici pripada, i svaka sadrži svoje osobine dobijene iz faze pretprocesiranja gde je bio upotrebljen Stanza paket, kao i da li je reč negativna ili ne.

5.1. Skup podataka

Skup podataka koji je korišćen u radu je SFU Review korpus koji se sastoji iz više manjih korpusa sa recenzijama raznih proizvođača. Za svaku kategoriju postoji 25 pozitivnih i 25 negativnih txt fajlova. Postoji i anotirana verzija ovog skupa podataka koja sadrži anotaciju negacije i spekulacije. Međutim za potrebe ovog rada vršena je zasebna anotacija. Skup rečenica je preuzet sa kaggle¹ sajta gde je isti ovaj korpus podeljen na pojedinačne rečenice i nalazi se u csv formatu. Zbog preobimnosti od 17 000 rečenica izdvojeno 2000.

5.2. Eksperiment 1 - detektovanje negativnih reči

U prvom eksperimentu korišćeni su atributi koji su navedeni u podsystemu za detekciju negativnih reči u poglavlju metodologija. Za obučavanje je korišćen SVM algoritam, gde je za kernel funkciju izabran RBF. Testiran je hiperparametar C, čije su vrednosti uzete u opsegu 0.01-0.1. Sistem je evaluiran pomoću kros-validacije i to je uzeta u obzir F1 mera. Bazni model je kreiran samo pomoću tokena i leme reči. Pokazano je da se najbolji rezultat dobija kada je vrednost C=0.1.

^[1] <https://www.kaggle.com/datasets/ma7555/the-sfu-review-corpus-negation-annotated>

U narednoj tabeli biće prikazani rezultati dobijeni raznim kombinacijama atributa.

Tabela 3. Razne kombinacije atributa za SVM algoritam

token, lema, pos	91.2%
token,lema,pos,bigram leme prethodne i posmatrane reči, pos prethodne reči	90.8%
token,lema,pos, bigram leme naredne i posmatrane reči, pos naredne reči	90.8%
token,lema,pos, bigram lema prethodne/naredne reči i posmatrane reči, pos prethodne/naredne reči	91.3%
token,lema,pos, bigram lema prethodne/naredne reči i posmatrane reči	91.35%
token,lema,pos, bigram lema prethodne/naredne reči i posmatrane reči, afiks, karakter 5-gram	91.44%

5.3. Eksperiment 2 - detektovanje opsega negacije

U drugom eksperimentu korišćeni su atributi koji su navedeni u podsystemu za detekciju opsega negacije u poglavlju metodologija. Testirane su razne kombinacije ovih atributa. Vršeno je BIO labelovanje gde je svaka negativna reč označena sa B, reči koje upadaju u opseg negacije sa I, a sve ostale reči u rečenici sa O. Opseg negacije se proteže u rasponu od 6 reči sa desne strane negativne reči i 3 reči sa leve strane. Za obučavanje je korišćen CRF, a u okviru njega lbfgs algoritam za obučavanje. Hiperparametri čije su razne vrednosti testirane su maksimalan broj iteracija (u opsegu 50-250), c1 i c2 koeficijenti u opsegu 0.1-0.3. Skup podataka je podeljen na obučavajući i test skup i korišćena je random vrsta uzorkovanja. Rešenja su evaluirana pomoću makro F1 mere pošto je u pitanju višeklasna klasifikacija. Najbolja kombinacija hiperparametara je sledeća: max_iterations=100, c1=0.3, c2=0.1, gde F1 mera iznosi 78.46%. Kombinacija atributa se nalaze u narednoj tabeli.

Tabela 4. Razne kombinacije atributa za CRF algoritam

token,lema,pos	77.02%
token,lema,pos,pos glave	68.43%
token,lema,pos, pos glave, deprel	71.54%
token,lema,pos, deprel	75.86%
token,lema,pos, deprel, bigram lema naredne i posmatrane reči	77.26%
token,lema,pos, deprel, bigram lema prethodne i posmatrane reči	80.65%
token,lema,pos, deprel, bigram lema naredne/prethodne i posmatrane reči	82.34%
token,lema,pos, deprel, bigram lema naredne/prethodne i posmatrane reči, usmerena zavisna putanja	83.82%

6. DISKUSIJA

Rezultati dobijeni u evaluaciji sistema su veoma dobri kako za detekciju negativnih reči (F1=91.44%), tako i za detekciju opsega negacije (83.82%). U radu [10] za detekciju negativnih reči rezultat iznosi 93.49%, a za opseg 77.98%. Na prvi pogled sistem za detektovanje negativnih reči je bolji u pomenutom radu, dok je detekcija opsega bolja u ovom radu. Međutim, treba uzeti u obzir više stvari.

U radu [10] detekcija negacije vršena je na ConanDoyle skupu podataka koji sadrži 5520 rečenica, dok u ovom radu skup sadrži 2000 rečenica. Iako sam postupak testiranja nije bio u potpunosti identičan, izveden je sličan zaključak, a to je da kombinacija tokena, leme, POS taga, bigrama posmatrane reči sa prethodnom/narednom reči, afiksa i 5-grama bez afiksa afikslnih negativnih reči daje najbolje rezultate.

Što se tiče opsega negacije, atributi takođe nisu bili identični. U radu [10] su se, pored ovde korišćenih atributa, koristili i POS tag glave drugog reda (glava glave), graf zavisnosti od svake reči do negativne reči, tip negacije (jednorečna, višerečna, afikslna). Sa druge strane, u pomenutom radu skup podataka je već bio anotiran sa negacijom, dok je u ovom radu anotacija poluautomatska i vršena je zasebno. Za opseg negacije nije uzeto u obzir kada se afikslna negativna reč nalazi na kraju rečenice pa ceo deo rečenice pre negativne reči spada u opseg negacije.

7. ZAKLJUČAK

U ovom radu predstavljen je sistem za anotaciju i detekciju negacije u rečenicama. Motivacija je bila to što postoje razni tekstovi na internetu čija analiza je od velikog značaja. Detekcija negacije je ključna u toj analizi. Predstavljen sistem sastoji se od pod sistema za pretprocesiranje teksta gde se vrši anotacija negativnih reči, pod sistema za detekciju negativnih reči i pod sistema za detekciju opsega negacije. Skup podataka nad kojim je testiran ovaj sistem sastoji se od 2000 rečenica iz SFU Review korpusa. Detekcija negativnih reči vršena je pomoću SVM algoritma, a opsega negacije pomoću CRF-a. Za evaluaciju SVM algoritma upotrebljena je krosvalidacija zasnovana na F1 meri i najbolji rezultat iznosi 91.44%. Za CRF korišćena je makro F1 mera i rezultat nad test skupom iznosi 83.82%.

U nastavku razvoja ovog sistema postoje razne stavke koje bi se mogle poboljšati. Pre svega, prilikom anotacije negativnih reči mogle bi se uzeti u obzir i negacije izražene sa više reči. Za određivanje opsega negacije treba ispitati i slučajeve kada je negativna afikslna reč na kraju rečenice. Što se tiče algoritama klasifikaciju, postoje i drugi atributi koji bi se mogli uzeti u obzir, kao na primer graf zavisnosti od svake reči do negativne reči. Algoritam koji je takođe veoma zastupljen u detekciji negacije jeste RNN neuronska mreža, pa bi se i ona mogla uključiti u sistem.

8. LITERATURA

- [1] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *J. Biomed. Informat.*, vol. 34, no. 5, 2001.
- [2] P. D. Turney and M. L. Littman, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus," 2002. Доступно: <https://arxiv.org/ftp/cs/papers/0212/0212012.pdf>
- [3] V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik, "The bioscope corpus: Biomedical texts annotated for uncertainty, negation and their scopes," *BMC Bioinf.*, vol. 9, no. 11, pp. 1–9, 2008.
- [4] S. Sohn, S. Wu, and C. G. Chute, "Dependency parser-based negation detection in clinical narratives," *AMIA Summits Transl. Sci. Proc.*, vol. 2012, Oct. 2012.
- [5] R. Morante and W. Daelemans, "Conandoyle-neg: Annotation of negation in conan doyle stories," in *Proc. 8th Int. Conf. Lang. Resour. Eval., Istanbul, Turkey, 2012*.
- [6] A. Hogenboom, P. van Iterson, B. Heerschoop, F. Frasincaar, and U. Kaymak, "Determining negation scope and strength in sentiment analysis," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2011.
- [7] I. Councill, R. McDonald, and L. Velikovich, "What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis," in *Proc. Workshop Negation Speculation Natural Lang. Process.*, 2010.
- [8] J. Reitan, J. Faret, B. Gambäck, and L. Bungum, "Negation scope detection for Twitter sentiment analysis," in *Proc. 6th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2015.
- [9] M. Enger, E. Velldal, and L. Øvrelid, "An open-source tool for negation detection: A maximum-margin approach," in *Proc. Workshop Comput. Semantics Beyond Events Roles, 2017*.
- [10] M. Enger, "A portable toolkit for detecting negation," in *UiO:Department Of Informatics, 2016*, pp.5-6.

Kratka biografija:



Aleksandra Milijević rođena je u Beogradu 1998. god. Master rad na Fakultetu tehničkih nauka iz oblasti Računarstva i automatike odbranila je 2022.god.

kontakt: aleksandramilijevic98@gmail.com