UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Ljubica Peleksić

KLASIFIKACIJA OBOLELIH OD ALCHAJMEROVE BOLESTI NA OSNOVU ANALIZE SPONTANOG GOVORA

master rad

Beograd, 2021.

Mentor:

doc. dr Jelena Graovac, profesor
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

prof. dr Gordana Pavlović-Lažetić

doc. dr Jovana Kovačević

Datum odbrane:

porodici

Sadržaj

Glava 1

Uvod

U današnjem svetu čovek ima mogućnost da veliki broj procesa automatizuje korišćenjem računara. Kako u drugim životnim oblastima, ovo se ispoljava i u medicini. Alchajmerova bolest je tip bolesti mozga koja sa vremenom postaje sve teža. Simptomi se dešavaju kada su neuroni u delovima mozga koji su zaduženi za učenje i memoriju, tj. kognitivne funkcije oštećeni. [1]. Prvi problemi koji se pojavljuju kod obolelih su poteskoće u govoru, kao i gubitak memorije. Kasnije u toku bolesti se pojavljuju i problemi pri obavljaju nekih jednostavih zadataka. Svake tri sekunde u svetu neko razvije demenciju, a Alchajmerova bolest je najčešci oblik demencije. Procenjuje se da danas oko 50 miliona ljudi u svetu imaju Alchajmerovu bolest ili demenciju vezanu za nju. Pretpostavlja se da bi ovaj broj mogao da naraste do 82 miliona do 2030. godine. i do 152 miliona do 2050. godine. [3].

Metodi za dijagnostifikovanje bolesti se sastoje iz struktuiranih intervjua koje izvode lekari. Tokom ovih intervjua je teško da se uoči kompleksna priroda nedostataka koji se mogu naći u govoru obolele osobe. Ovi intervjui testiraju jezičke sposobnosti i uključuju imenovanje objekata, izgovaranje jedne reči ili generisanje reči iz datog konteksta[2]. Dijagnostifikovanje Alchajmerove bolesti je najlakše iz ugla porodice i prijatelja, pošto su oni u stanju da u svakodnevnom životu, pri normalim konverazijama, primete male promene u ponašanju bližnjih.

Teži se automatizovanom pristupu sa objektivnim metodama dijanostifikovanja bolesti, kao i određivanja stepena progresije. Ovakav novi pristup je neophodan kako zbog preciznosti i zbog ranog dijagnostifikovanja bolesti, tako i zbog brzine dobijanja dijagnoze i činjenice da bi time dijanozu i tretman za ublažavanje simptoma moglo dobiti više ljudi u svetu, medju kojima danas mnogi nemaju pristup lekarima i lekarskoj nezi.

Prema Svetskom Alchamjerovom Izveštaju iz 2011. godine postoje benefiti rane dijagnoze i intervencije koji su efektivniji u ranim stadijumima bolesti i omogućava osobama da isplaniraju negu i tretmane koje žele da imaju u budućnosti. Pored korišćenja lekova, a ima ih nekoliko, smanjenje pogoršanja u kognitivnih funkcija osobe može pružiti i specijalizovana terapija.

U više naučnih radova i studija je pokazano da korišćenjem obrade prirodnog jezika (eng. natural language processing) i mašiskog učenja (eng. machine learning) se može doći do vrednih podataka. Motivacija za ovom vrstom obrade podataka i zaključivanja, koji će biti predstavljeni u narednim poglavljima, je nastala iz [2] i [5]

Problem koji se rešava u okviru ovog rada je da li je moguce primeniti automatizaciju postupka dijagnostifikovanja obolelih od Alchajmerove bolesti dovoljno precizno, tačnije da li je moguće razlikovati starije subjekte od subjekata koji pate od Alchajmerove bolesti na osnovu spontanog govora. Ono što pojedinac izgovara je nepresušan izvor informacija o njemu samom, a upravo tom činjenicom se vodimo kada vidimo mogućnost u nalaženju dijanoze analizom spontanog govora. Dva pravca istraživanja su pracena, jedan leksički i drugi putem metoda mašinskog učenja. U sklopu leksičkog pristupa, odredjene, vec poznate metrike se izračunavaju za svaki intervju, koji je predstavljen skupom reči, njenom vrstom reči i lemom. U okviru ovog pristupa tražimo pravilnosti u govoru koje nam mogu pokazati da osoba boluje baš od ove bolesti. U pristupu mašinskim učenjem ne postoje poznate metrike, već se prepušta algoritmima da uoče zakonitosti u podacima u obliku učestalosti korišćenja reči i na taj način se preciziraju zakonitosti i razlike izmedju dve grupe pacijenata.

Do danas nije pronadjen lek koji zaustavlja ili usporava razvoj bolesti, ali postoje lekovi za koje se smatra da mogu da uspore bolest ako je rano dijagnostifikovana, kao i oni za koje se smatra da poboljšavaju kognitivne funkcije obolele osobe.[1].

Ovu temu razradjujemo kroz šest poglavlja, gde se nakon uvodnog poglavlja osvrćemo na podatke korišćene u praktičnom radu, način sakupljanja i obrade podataka. U drugom poglavlju je opisan način rešavanja ovog problema metodama leksičke analize, gde su predstavljene metrike koje su korišćene kao i motivacija za njihov izbor. Treće poglavlje predstavlja prikaz rešavanja problema metodama mašinskog učenja, modele reprezentacije teksta kao i algoritme mašinskog učenja koji su iskorišćeni u praktičnom radu. U okviru petog poglavlja se diskutuju dobijeni rezultati reprezentativnog uzorka, primenama svih metoda, dok se u poslednjem poglavlju, u okviru zaključka, sumiraju dobijeni rezultati i korišćene metode.

Glava 2

Podaci

U ovom poglavlju će biti predstavljeni podaci koji su korišćeni u radu za dobijanje rezultata, kao i tehnike obrade podataka koje su vršene nad njima. Podaci korišćeni za rešavanje problema klasifikacije obolelih osoba od Alchajmerove bolesti su transkripti slobodnog govora dementnih osoba. Transkripti se prikupljaju kao audio ili video zapisi, a razgovori sa osobama nisu struktuirani niti imaju neki odredjeni sled. Na osobi koja intervjuiše je da postavlja pitanja, dok intervjuisani odgovara. Osoba koja intervjuiše takodje prati tok razgovora sa pitanjima, te nisu u svakom razgovoru ista. U svrhu poredjenja, pored podataka o osobama obolelim od demencije tipa Alchajmer, bili su potrebni i intervjui sa nedementnim starijim osobama. Intervjui sa osobama obolelim od Alchajmera su sakupljani u obliku video i audio zapisa koji su prikupljeni od osoba koje su koristile usluge dnevnog boravka za obolele od Alchajmera u Novom Sadu, jedine takve ustanove u Srbiji organizovane od strane Udruženja građana Alchajmer. Ovi intervjui su bili prikpljani tokom grupnih razgovora sa obolelim. Intervjui sa nedementnim starijim osobama su prikupljeni zahvaljujući aktivnostima više od 20 volontera studenata Biološkog fakulteta u Beogradu, krajem 2017. godine i u toku 2018. Od audio zapisa razgovora sa osobama obolelim od Alchajmera su kreirani transkripti, dok su studenti koji su učestvovali u prikupljanju podataka od dementnih starijih osoba zatim od audio zapisa kreirali tekstualne zapise. Svaki intervju je zapisan na dva načina, kao originalan, u kome je svaka reč napisana tačno onako kako je izgovorena, uključujući ponavljanja, nedovršene reči, greške u izgovoru i drugi u kome su ispravljene sve greške u izgovoru. Kako su ovi intervjui sprovedeni grupno, prvi korak je bio da se razdvoje razgovori, tako da rečenice koje je izgovorila jedna osoba se nalaze u jednoj tekstualnoj datoteci koja nosi ime te osobe. Reči osobe koja postavlja pitanja

se ne moraju naći u svakoj datoteci i ne predstavljaju podatke koji se obradjuju. Nakon toga transkripti su pažljivo obradjeni, tako da svaki bude ispravno podeljen na rečenice. Ovaj korak je bio veoma važan kako bi u narednom koraku moglo da se dodje do vrsta reči i lema, za svaku izgovorenu reč.

Transkripti su zapisivani po definisanom protokolu koji ubraja sledeće:

1. Intervju sa jednom osobom se nalazi u datoteci koja nosi ime te osobe
2. Ime obolelog se označava vitičastim zagradama
3. Pitanje postavljeno od strane osobe koja vodi intervju se označava uglastim zagradama
4. Koristi se UFT-8 kodna šema i latinica
5. Koriste se slova sa dijakriticima (č, ć, š, đ,...)
6. Pauze između izgovorenih reči se zapisuju odgovarajućim brojem crtica, gde svaka crtica predstavlja jedan sekund pauze
7. Brojevi se zapisuju sa crticom izmedju, ako su višecifreni brojevi

Nakon što su transkripti bili ispravno zapisani po protokolu, a rečenice ispravno podeljene, takve datoteke su prosleđene na Filološki fakultet u Beogradu, gde su algoritamskim putem za svaku reč odradjeni vrsta reči i njena lema. Vrste reči biće upotrebljene u okviru rešavanja problema klasifikacije obolelih pacijenata od Alchajmerove bolesti leksičkim metodama. Lema reči se koristi pri računanju vrednosti vokabulara i vokabulara za reči izgovorene jednom. Ove metrike će biti pomenute u daljem radu, u sekciji koja se bavi rešavanjem problema metodama leksičke analize.

2.1 Problem odredjivanja vrsta reči

U daljem tekstu će biti objašnjen problem dodeljivanja vrsta reči sekvenci reči, kao i predstavljena dva algoritma za rešavanje ovog problema:

1. Skriveni Markovljev Model (eng. Hidden Markov Model / HMM)
2. Uslovljena nasumična polja (eng. Conditional Random Fields/CRF)

Vrste reči koje postoje i njigove oznake na engleskom jeziku su sledece:

1. Pridev: ADJ
2. Apozicija: ADP
3. Prilog: ADV
4. Pomoćni: AUX
5. Veznici: CCONJ
6. Član: DET
7. Uzvik: INTJ
8. Imenica: NOUN
9. Broj: NUM
10. Rečca: PART
11. Zamenica: PRON
12. Vlastita imenica: PROPN
13. Znak interpunkcije: PUNCT
14. Veznik: SCONJ
15. Simbol: SYM
16. Glagol: VERB
17. Drugo: X

Metoda odredjivnja vrsta reči je kompleksan problem. Zadatak je takav da dodeljujemo svakoj reči xi u ulaznoj sekvenci reči labelu yi, tako da izlazna sekvenca Y ima istu dužinu kao ulazna sekvenca X. [4]. Reči imaju vise mogućih vrsta reči i zadatak ovog procesa je da se pronadje ispravna vrsta reči za svaku pojedinu situaciju. Za reči se određuje takva vrsta koja je najverovatnija. Za mnoge je verovatnoća da pripadaju svim vrstama osim jednoj izuzetno mala, pa je lako odlučiti se.

Jedan od načina da se reši problem odredjivanja vrsta reči je da se koristi Skriveni Markovljev model (eng. Hidden Markov Model / HMM). Skriveni Markovljev

model je probabilistički sekvencni model koji za sekvencu jedinica izračunava raspodelu verovatnoće po mogućim sekvencama i bira najbolju opciju. Markovljev lanac, model koji nam govori o verovatnoćama sekvenci slučajnih promenljivih, ima pretpostavku da ako želimo da predvidimo buduće stanje, jedino što je bitno je trenutno stanje. Markovljev lanac se grafički predstavlja grafom, gde su čvorovi grafa stanja, a grane predstavljaju verovatnoće. Suma svih vrednosti grana koje idu iz odredjenog čvora mora biti jedan. Markovljev lanac se oslanja na dogadjaje koji mogu da se posmatraju, dok u slučaju reči to nije moguće. Zato se koriste Skriveni Markovljev Model., koji poseduje skrivene promenljive. Zadatak odredjivanja skrivene sekvence promenljivih na osnovu observacija u modelu se naziva dekodiranje. Skriveni Markovljev Model počiva na dve pretpostavke. Prva je identična kao za Markovljev lanac, dok druga kaže da verovatnoća nekog stanja zavisi samo od stanja koje je proizvelo to stanje i ni jednog više. Algoritam za odredjivanje vrsta reči se sastaji iz matrice koja sadrži verovatnoće da se jedna vrsta reči nalazi posle druge i matrice koja sadrži verovatnoće da se odredjena vrsta dodeli odredjenoj reči. Algoritam za dekodiranje za Skriveni Markovljev Model se naziva Viterbi algoritam. Viterbi algoritam prima dve matrice koje smo pomenuli, a vraća putanju kroz stanja Skrivenog Markovljevog Modela koja deeljuje najveću verovatnoću datoj sekvenci. [4].

Algoritam koji je prethodno prikazan ima problem sa nepoznatim rečima, vlastitim imenima, akronimima, novim rečima. Algoritam uslovljenih nasumčnih polja nalazi način da iskoristi odredjene odlike reči, kao što su veliko slovo ili prefiks ili sufiks reči, što je teško dodati u Skriveni Markovljev Model. Trenira se logaritamski linearan model. U modelu uslovljenih polja računamo verovatnoću svih vrsta reči u sekvenci, a ne pojedinačni vrstu jedne po jedne reči. Svako svojstvo se oslanja na vrstu reči prethodne i sledeće reči i na celu ulaznu sekvencu reči. Za zaključivanje se takodje koristi Viterbi algoritam da bi se odabrala najbolja sekvenca vrsta reči. [4].

2.2 Lematizacija

Proces lematizacije jeste onaj u kome se za svaku reč nalazi njena kanonska forma, tačnije lema. U srpskom jeziku, za imenice lema je nominativ jednine, za glagole infinitiv, a za prideve nominativ jednine muškog roda. Takođe, proces lematizacije uključuje vraćanje rodne varijacije reči na njen pređašnji oblik. Postoje dva metoda za rešavanje problema lematizacije. Prvi da se prema obliku reči otklanja sufiks i

nalazi njena lema. Ovaj oblik rešavanja ne daje tako dobre rezultate. Drugi pristup uključuje korišćenje skupa podataka koji za svaku reč, za svaku njenu moguću vrstu reči, ima odredjenu lemu. Postoji mogućnost i kombinovanja ova dva pristupa.

2.3 Skup podataka pre i nekon obrade

Skup podataka koji će biti korišćen za eksperimente u ovom radu sadrži sakupljene podatke na već opisan način, koji su zatim obradjeni pomenutim tehinikama. Intervjua sa pacijentima obolelim od demencije Alchajmerovog tipa ima 22, koje nazivamo „pozitivni" i nalaze se u fascikli „P". Intervjua sa starijim licima koji nemaju utvrdjenu demenciju Alchajmerovog tipa ima 57 i oni se nalaze u fascikli „N", a nazivamo ih „negativni". Nakon procesa odredjivanja vrsta reči i lema za svaku reč intervjua, za svaki transkript je kreiran novi dokument koji u sebi sadrzi potrebne informacije. Ako se u transkriptu nalazi rečenica:

Ja sam Bojana.

Onda će se u odgovarajućoj datoteci naći i sledeći redovi, gde u svakom prva reč označava izvorni oblik reči, druga vrstu, a treća lemu.

Ja ADV Ja sam AUX jesam Bojana PROPN Bojana . PUNCT .

Glava 3

Rešavanje problema metodama leksičke analize

Glava 4

Rešavanje problema metodama mašinskog učenja

4.1 Vektorska reprezentacija teksta

4.2 Vreća reči

4.3 N-grami

4.4 TF metrika

4.5 TF-IDF metrika

Glava 5

Rezultati

Glava 6

Zaključak

Bibliografija

[1] Alzheimer's Association. Alzheimer disease facts and figures. In *Speech and Language Processing*, 2021.

[2] Vlado Kešelj Calvin Thomas and Elissa Asp Nick Cercone, Kenneth Rockwood. Automatic detection and rating of demetia of alzheimer type through lexical analysis of spontaneous speech.

[3] Alzheimer's Disease International. Dementia statistics. 2020.

[4] Daniel Jurafsky and James H. Martin. Sequence labeling for parts of speech and named entities. In *Speech and Language Processing*, 2020.

[5] Jed A. Meltzer Kathleen C. Fraser and Frank Rudzicz. Linguistic feature identify alzheimer's disease in narrative speech.

Biografija autora

Ljubica Peleksić (*Beograd, 18. novembar 1993.*) Ljubicina biografija