Obrada prirodnih jezika

Projekat za školsku 2021/2022. godinu

Tema projekta

Tema predmetnog projekta za školsku 2021/2022 godinu se tiče problema aspektne analize sentimenta tekstova na srpskom jeziku. Cilj je da se napravi sistem koji na osnovu zadatog kratkog teksta (dužine od nekoliko reči do par rečenica) vraća aspektne kategorije prema kojima je u tekstu izražen sentiment, kao i oznaku sentimenta (pozitivan, negativan ili neutralan) za svaku aspektnu kategoriju. Na primer, ako se među aspektnim kategorijama za domen hotela nalaze *lokacija* i *cena*, za iskaz:

„Hotel jeste dosta udaljen od plaže, ali je soba bila vrlo povoljna.“

sistem bi trebalo da prepozna da se u tekstu izražava negativan sentiment prema aspektnoj kategoriji *lokacija*, a pozitivan prema aspektnoj kategoriji *cena*.

Projekti će se izrađivati grupno. Proces prijavljivanja grupe je opisan u odeljku o propozicijama izrade projekta. Projekat se može implementirati u programskom jeziku i paketu po izboru.

Definisanje aspektnih kategorija znatno zavisi od tematskog domena tekstova u pitanju, te je jedan od početnih ciljeva u ovom zadatku formulisanje adekvatnog skupa aspektnih kategorija za odabrani domen tekstova. Svaka projektna grupa će razmatrati drugačiji tematski domen tekstova (npr. domen hotela, domen mobilnih telefona, domen video igara, itd.).

Izrada projekta podrazumeva prikupljanje odgovarajućeg skupa polaznih tekstova na srpskom jeziku iz odabranog domena, kao i ručnu anotaciju aspektnih kategorija i sentimenata koji su izraženi prema svakom od aspekata. Tako kreirani skup podataka je zatim potrebno iskoristiti za obučavanje i evaluaciju nekoliko različitih statističkih modela.

Neophodno je učešće svih članova grupe u svim fazama izrade projekta, tj. nije dozvoljena podela posla između članova grupe po fazama. U nastavku će biti detaljnije opisana svaka od faza.

Faza 1 - Prikupljanje podataka

Proces prikupljanja podataka podrazumeva formiranje dovoljno velikog skupa tekstova za odabrani tematski domen, tako da se takav skup može nakon anotacije iskoristiti za obučavanje i evaluaciju statističkih modela. Stoga pri izboru tematskog domena treba najpre razmotriti za koji domen je moguće pronaći dovoljne količine tekstova na srpskom. Kao izvor podataka za formiranje ovakvog skupa tekstova mogu poslužiti bilo koji javno dostupni veb sajtovi sa sadržajem na srpskom jeziku.

Formirani skup treba da sadrži minimalan broj tekstova koji je određen veličinom grupe. Prikupljanje (i anotiranje) većeg broja tekstova od navedenog minimuma donosi dodatne bodove pri određivanju ocene. Dozvoljeno je da se u tekstovima pisanim na srpskom jeziku javi i poneki termin napisan na engleskom, pri čemu je za očekivati da će ovo u nekim tematskim domenima (npr. onim vezanim za

tehničke uređaje) biti frekventnija pojava nego u drugim. Finalni skup prikupljenih tekstova treba pročistiti od dupliranih unosa istog teksta.

Prikupljeni tekstovi treba da budu sačuvani u vidu jednog tab-separated UTF-8 enkodovanog TXT fajla, gde svaki red fajla sadrži tri kolone sledećeg sadržaja:

1. *Domain* – jedinstveni identifikator tematskog domena iz koga su tekstovi prikupljeni (slobodno definisan od strane grupe)
2. *Source* – URL ili jedinstveni identifikator izvora iz koga je tekst dobijen
3. *Raw Text* – prikupljeni kratak tekst u izvornom obliku

Faza 2 – Anotacija podataka

U ovoj fazi potrebno je u svakom od prikupljenih tekstova ručno obeležiti aspektne kategorije prema kojima je izražen sentiment, kao i označiti koji tip sentimenta je u pitanju - pozitivan, negativan ili neutralan. Za sentiment treba koristiti oznake POS, NEG ili NEUT. Aspektne kategorije će se prirodno znatno razlikovati od domena do domena, te stoga njihove oznake mogu biti slobodno definisane od strane grupe. Očekuje se da skup aspektnih kategorija bude što potpuniji, tj. da pokrije sve relevantne stavke prema kojima se u tekstovima izražava sentiment. U definisanju skupa aspektnih kategorija za određeni domen, dozvoljeno je konsultovanje relevantnih naučnih radova, ali se takođe očekuje i uvid u konkretne prikupljene podatke, da bi se osigurala potpunost. Pri tome, za očekivati je da aspektne kategorije objedinjuju širi skup aspektnih izraza (npr. kategorija *Hrana* u domenu restorana može obuhvatiti veliki broj različitih naziva konkretnih jela).

Za sprovođenje anotacije dozvoljeno je, ali ne i neophodno, koristiti bilo koji namenski program za anotaciju, bilo neki postojeći, bilo neki razvijen od strane grupe. U sprovođenju anotacije treba pratiti standardnu metodologiju označavanja podataka, koja podrazumeva:

1. Odabir oznaka tj. definisanje skupa aspektnih kategorija
2. Formulisanje uputstava za anotaciju – ova uputstva bi trebalo da sadrže jasne definicije za svaku od oznaka koje se koriste u anotaciji (npr. šta sve spada u određenu aspektnu kategoriju), kao i usaglašene instrukcije za sistematsko postupanje u karakterističnim problematičnim situacijama (npr. kada neke karakteristične izraze ne treba tretirati kao neutralne nego kao pozitivne/negativne ili obratno)
3. Kalibraciju – proveru upotrebljivosti kreiranog skupa oznaka i uputstava za anotaciju uz pomoć malog podskupa primera tekstova (oko 10% od ukupnog broja) koje svi članovi grupe treba da paralelno anotiraju, zasebno i bez međusobnih konsultacija. Ako se u ovom koraku uoče nedostaci u skupu oznaka ili u uputstvima, treba se vratiti na neki od prethodnih koraka.
4. Sprovođenje anotacije – podatke bi trebalo ravnomerno rasporediti između svih članova grupe, tako da svako anotira približno istu količinu podataka. Očekuje se da anotacija glavnog seta tekstova bude jednostruka, ali nije zabranjeno višestruko paralelno označavanje, ako članovi grupe procene da se time znatno podiže konzistentnost generisanih oznaka.
5. Analizu anotacije – određivanje saglasnosti anotatora na osnovu kalibracionog skupa (procentualan stepen saglasnosti između svaka dva člana grupe, kao i grupni prosek binarnih stepena saglasnosti), statističku analizu oznaka u finalnim podacima, itd.

Anotirane podatke treba sačuvati u vidu proširenja tab-separated UTF-8 TXT fajla iz prethodne faze, sa dodatom kolonom za anotacije u sledećem obliku:

(ASPEKT1, SENTIMENT1); (ASPEKT2, SENTIMENT2);…

Na primer, za tekst iz hotelskog domena naveden na početku postavke projekta, anotacija bi bila:

(lokacija, NEG); (cena, POS)

Skup za kalibraciju treba sačuvati u vidu posebnog tab-separated UTF-8 TXT fajla, koji bi sadržao samo 10% tekstova koji su korišćeni pri kalibraciji. Ovaj fajl treba da ima onoliko dodatih kolona za anotaciju koliko ima članova grupe, pri čemu u svakoj koloni treba koristiti isti format zapisa anotacija kao i u glavnom fajlu sa svim anotiranim tekstovima.

Faza 3 - Obučavanje i evaluacija statističkih modela

U ovoj fazi potrebno je razviti statističke modele za rešavanje tri klasifikaciona problema:

1. Detekcija aspektnih kategorija u zadatom tekstu – tekst se može odnositi na jednu, više, ili nijednu od aspektnih kategorija. Svakoj od aspektnih kategorija treba dodeliti po jedan binarni klasifikator koji predviđa da li se zadati tekst odnosi na njegovu aspektnu kategoriju.
2. Detekcija sentimenta za određenu aspektnu kategoriju u zadatom tekstu – za određenu aspektnu kategoriju koja je prisutna u zadatom tekstu potrebno je vratiti da li se prema njoj izražava pozitivan, negativan ili neutralan sentiment. Svakoj od aspektnih kategorija treba dodeliti po jedan višeklasni klasifikator koji predviđa sentiment (POS/NEG/NEUT) koji je izražen prema njegovoj aspektnoj kategoriji.
3. Zajednička detekcija aspektnih kategorija i sentimenata izraženih prema njima, kroz korišćenje posebnog višeklasnog klasifikatora za svaku aspektnu kategoriju, koji predviđa da li se zadati tekst odnosi na njegovu aspektnu kategoriju sa određenim sentimentom (POS/NEG/NEUT/NONE). Takođe je potrebno razmotriti da li ovaj pristup daje bolje rezultate u predikciji parova (aspekt, sentiment) od kombinovanja pristupa 1 i 2.

U pristupe koje je obavezno potrebno razmotriti spadaju tri klasifikatora predstavljena na predavanjima – multinomijalni naivni Bajesov klasifikator, logistička regresija i metoda potpornih vektora – korišćeni u kombinaciji sa odlikama dobijenim po principu vreće reči/n-grama. Za sve klasifikatore treba ispitati efekte različitih tehnika pretprocesiranja teksta. Počevši od osnovnih *bag-of-words* podešavanja gde se ne koriste sledeće tehnike, sistematski razmotriti sledeće:

- Normalizaciju svih tekstova na mala slova (lowercasing)
- Binarizaciju vrednosti *bag-of-words* odlika
- Frekvencijsko filtriranje reči
- TF, IDF i TFIDF ponderisanje
- Filtriranje stop-reči i/ili stemovanje reči (po izboru)
- Korišćenje bigrama i trigrama

Kao listu stop-reči moguće je koristiti neku od javno dostupnih lista ili formirati sopstvenu, pri čemu odabir određenog skupa stop-reči mora biti opravdan za konkretan domen tekstova koji se razmatraju. Za stemovanje reči na srpskom koristiti stemer Ljubešića i Pandžića iz paketa *SCStemmers* (dozvoljeno je i korišćenje alternativnih implementacija ovog stemera).

Obučavanje i evaluaciju modela je potrebno sprovesti putem 10-slojne stratifikovane unakrsne validacije, korišćenjem odgovarajuće metrike za merenje performansi. Pri tome je kod logističke

regresije i metode potpornih vektora potrebno sprovesti optimizaciju hiperparametra C / lambda koji određuje jačinu regularizacije, korišćenjem ugnežđene unakrsne validacije. Inicijalnim ispitivanjem, korišćenjem default vrednosti za ostala podešavanja, treba utvrditi koja od varijanti funkcije regularizacije (L1 / L2) i funkcije gubitka kod metode potpornih vektora (L1 / L2) daje bolje rezultate – ako nema primetnih razlika, preporučljivo je koristiti L2 oblike funkcija.

U naprednije pristupe čije razmatranje donosi dodatne bodove pri određivanju ocene spada korišćenje drugačijih ili dodatnih odlika u klasifikaciji, ili korišćenje drugačijih klasifikacionih modela ili konceptualnih pristupa zadatku. Ovde će biti navedeno nekoliko mogućih ideja, ali su studenti slobodni da u dogovoru sa predavačem razmotre i drugačije naprednije tehnike:

- Korišćenje vektora značenja reči (*word embeddings*) kao (dodatnih) odlika
- Korišćenje sentiment leksikona, u vidu (dodatnih) odlika i u vidu konceptualnog pristupa zadatku bez algoritama mašinskog učenja
- Parsiranje tekstova i definisanje odlika na osnovu karakterističnih sintaktičkih struktura
- Parsiranje tekstova i određivanje aspektnih sentimenata u svakoj klauzi/rečenici zasebno
- Obrada negacija

Grupe koje se odluče za razmatranje naprednijih pristupa bi trebalo da u dogovoru sa predavačem detaljnije razmotre implementaciju odabranih tehnika.

Propozicije izrade projekta

Optimalna veličina grupe je četvoro studenata. Minimalan ukupan broj tekstova koje treba prikupiti za takve grupe je 3000, i bar po 1000 primera za svaku aspektnu kategoriju, pri čemu je očekivano da se većina tekstova odnosi na više od jedne aspektne kategorije. Dozvoljene su grupe i od troje ili petoro članova, u kojem slučaju minimalan broj tekstova koji treba prikupiti iznosi 2500, odnosno 3500, uz isti minimum od bar 1000 primera za svaku aspektnu kategoriju. Kako se zahtevi u pogledu obučavanja i evaluacije statističkih modela ne razlikuju u zavisnosti od veličine grupe, preporučuje se da se studenti organizuju u veće grupe.

Studenti se mogu sami organizovati u grupe, za šta je otvoren i poseban kanal u okviru predmetnog tima na Teams platformi. Pre otpočinjanja rada na projektu, neophodno je formirati i zvanično prijaviti grupu putem mejla, na adresu: vuk.batanovic@ic.etf.bg.ac.rs. Prilikom prijave grupe, neophodno je navesti spisak članova grupe, izbor tematskog domena tekstova koje bi grupa želela da razmatra, i spisak izvora podataka koji bi se koristili za taj domen. Grupi će zatim u najkraćem roku biti zvanično dodeljen tematski domen ako nije već zauzet od strane neke ranije prijavljene grupe.

Za slučaj nemogućnosti samoorganizovanja u grupe, studenti mogu i da se individualno prijave za izradu projekta. U tom slučaju, biće od strane predavača organizovani u grupe sa ostalim studentima koji su se individualno prijavili ili će, u slučaju nedovoljnog broja tako prijavljenih studenata, biti pridruženi nekoj od već formiranih grupa. U oba slučaja, individualno prijavljeni studenti neće imati mogućnost izbora tematskog domena tekstova koje razmatraju.

Ova postavka predmetnog projekta će važiti do prolećnog semestra naredne školske godine. Grupe je potrebno formirati i prijaviti najkasnije do 01.08.2022, bez obzira na konkretan ispitni rok u kome se planira odbrana. Individualne prijave treba dostaviti najkasnije do 01.07.2022. Ni grupne ni individualne prijave nakon tih datuma neće biti uzimane u obzir.

Grupe koje žele da brane projekat u određenom ispitnom roku treba da pošalju urađeno rešenje i projektnu dokumentaciju do početka istog ispitnog roka, na adresu: vuk.batanovic@ic.etf.bg.ac.rs.

U projektnoj dokumentaciji treba detaljno opisati svaku od faza izrade projekta. Ovo podrazumeva temeljno opisivanje procesa prikupljanja podataka, izvora podataka, navođenje kriterijuma koji su u tom procesu korišćeni i opisivanje kako je proces obavljen sa tehničkog aspekta. Pored toga, dokumentacija mora sadržati detaljan opis anotacije podataka, uključujući uputstva za anotaciju, kao i opis tehničke strane označavanja podataka. Takođe se očekuje da izveštaj sadrži deskriptivni statistički prikaz prikupljenih i anotiranih podataka. Za fazu obučavanja i evaluacije statističkih modela podrazumeva se da izveštaj sadrži pregledni tabelarni prikaz rezultata različitih modela i efekata različitih razmotrenih podešavanja. Dokumentacija ne treba da sadrži iskopirana detaljna objašnjenja iz nastavnih materijala za korišćene tehnike i algoritme.

Ukoliko su projektna rešenja i dokumentacija adekvatni, u dogovoru sa studentima biće određen termin odbrane projekta u toku ispitnog roka. Odbrane će biti moguće u svim ispitnim rokovima predviđenim za predmete iz letnjeg semestra, a održavaće se bilo uživo bilo preko Teams platforme, uzimajući u obzir epidemiološku situaciju u konkretnom roku i dogovor grupe sa predavačem.

Ocene će se dobijati na osnovu broja prikupljenih bodova na skali 0-100, prema sledećoj raspodeli:

- 25 poena – faza prikupljanja podataka
- 25 poena – faza anotacije podataka
- 25 poena – faza obučavanje i evaluacije statističkih modela
- 15 poena – kvalitet i potpunost priložene projektne dokumentacije; poštovanje projektne specifikacije očekivanog formata podataka
- 25 poena – rad iznad traženih minimalnih zahteva (veća količina prikupljenih i anotiranih podataka, razmatranje naprednijih statističkih pristupa/odlika, itd.)

Za svaku od prve četiri stavke neophodno je da grupa ostvari barem polovinu od minimalnog broja poena. Drugim rečima, nije moguće odbraniti projekat bez sprovođenja i opisivanja sve tri faze izrade. Za maksimalnu ocenu (10) potrebno je barem u nekom delu projekta nadmašiti postavljene minimalne zahteve. Pri tome, dozvoljeno je da se samo neki članovi grupe opredele za ovaj cilj, a da se ostatak grupe ocenjuje shodno minimalnim projektnim zahtevima.