Uspješnost altruističnih zahtjeva

Loredana Musap

Mara Šumelj

Prirodoslovno-matematički fakultet

Sveučilište u zagrebu

1. Uvod

U današnje vrijeme kad ljudi "žive" na internetu sve je raširenije traženje pomoći preko istog. Bilo koji problem na koji pojedinac naiđe okreće se "internet zajednici" za pomoć. Iz tog razloga postale su sve popularnije stranice kao što su DoonorsChoose.org, StackOverflow.com ili Reddit.com. Na takvim stranicama ljudi postavljaju pitanja, a dobar odgovor ovisi isključivo o drugim ljudima uključenim u te društvene mreže.

Stranica koju ćemo proučavati, Random Acts of pizza[1] (dio Reddita), zamišljena je na vrlo jednostavan način. Bilo koji član može napisati molbu za besplatnu pizzu s kratkom porukom ostalim čitateljima. Ako njegov zahtjev ostavi dobar dojam baš će on biti izabran od drugog člana koji nekoga želi usrećiti pizzom. Onaj tko želi pokloniti pizzu odabrat će molbu koja najviše odgovara njegovim motivima.

Kako pridobiti nekoga i dobiti bespatnu pizzu? Koji su to faktori koji najviše utječu na izbor?

Komunikacija je ključni element društvenih zajednica, pa je stoga uspjeh u komunikaciji jednako bitan za internet zajednicu kao i za pojedinca. Obzirom da posljednjih godina broj internet zajednica za raspravu, socijalnu podršku, razgovor, zabavu i sl. neprestano raste tako je pitanje uspjeha u komunikaciji postao sve bitniji i zanimljiviji. Članovima koji postavljaju pitanja na forumima bitno je da što brže dobiju odgovor, na socijalnim mrežama ljudi žele uspostaviti što bolji

[1] Random Acts of pizza - *www.reddit.com/r/Random_Acts_Of_Pizza*

kontakt s drugima i ostvariti što više prijateljstava, prilikom rasprava svi žele dobiti što više pohvala i komentara prilikom izražavanja svog mišljenja o nekoj temi te se tako povezati sa što više istomišljenika. Zašto su neki postovi uspješniji od drugih? Što točno potiče ljude da što više komentiraju i da se uključuju u rasprave? Da bi zajednice na internetu bile što posjećenije i komentiranije bitno je znati što pokreće ljude da odgovaraju na razne zahtjeve na internetu.

2. Opis problema

Problem kojim se mi bavimo u ovom radu predstavljen je na stranici Kaggle[2] te su nam tamo ponuđeni podatci koje ćemo koristiti. Podatci su preuzeti sa stranice Reddit i u prilagođenom formatu. Set podataka se sastoji od 4040 tekstualnih zahtjeva za pizzu koji su skupljani između 8.12.2010. i 29.9.2013. Svaki zahtjev osim tekstulanog dijela iz kojeg se metodama obrade prirodnog jezika mogu izvući dodatni atributi sadrži i već gotove atribute koji se odnose na društveni kontekst zahtjeva. Iako postoje mnogi atributi koji opisuju korisnika i sam zahtjev u trenutku preuzimanja mi smo se odlučile proučavati samo atribute koji se odnose na trenutak kada je zahtjev postavljen na stranicu. Smatramo da bi atributi koji su se mijenjali nakon postavljanja zahtjeva na stranicu mogli negativno utjecati na predikciju na još neviđenim zahtjevima.

[2] Kaggle problem - *www.kaggle.com/c/random-acts-of-pizza*

3. Opis metode

Odabir značajki jako je bitan, čak i krucijalan u ovom projektu. Obzirom da u podatcima imamo jako puno atributa mnogi su nebitni i mijenjaju rezultat u krivom smjeru ili ga ne mijenjaju uopće. Stoga smo metodom nadziranog učenja odabrali najbitnije a one loše smo izbacili.

3.1. Lingvistički faktori

Prvi korak rješavanja problema biti će dobro analizirati tekst same molbe, odnosno izvući odgovarajuće lingvističke atribute iz samog teksta. Dakle, one atribute koji točno karakteriziraju sam tekst i daju dobru informaciju o toj specifičnoj molbi. Da bi izvukli najviše korisnih informacija iz teksta važno je obaviti samo preprocesiranje teksta.

Prvo smo izbacili dijakritičke znakove (. ! ? , *) te su nam u svim zahtjevima ostale samo riječi. Međutim neke riječi imaju veću klasifikacijsku vrijednost od drugih. Riječi s malom klasifikacijskom vrijednosti je poželjno eliminirati. Obzirom da su naši tekstovi na engleskom jeziku da bi eliminirali nepoželjne riječi koristili smo listu neinformativnih riječi u engleskom jeziku, tj. stop-words-list.

Time smo za sve lingvističke faktore dobili podlogu. Razmatrajući zahtjeve došli smo do zaključka da su najvažniji faktori: evidentnost, recipročnost, sentimentalnost i duljina.

Evidentnost – neki zahtjevi potkrjepljuju svoju priču dokazima npr. slikama, linkovima, videima. Koristili smo najjednostavniju mjeru evidentnosti: prisutnost slike u tekstualnom zahtjevu.

Duljina – brojimo riječi u zahtjevu te pretpostavljamo da duljina teksta pokazuje trud koji je uložen pri pisanju.

Recipročnost – želimo svaki zahtjev ocijeniti kao pozitivan, negativan ili neutralan. Da bi odredili recipročnost u pojedinačnom zahtjevu pobrojali smo pozitivne i negativne riječi koje se pojavljuju te smo dobili varijablu $s = pozitivni - negativni$. Ako je s > 0 zahtjev je pozitivan, ako je s < 0 zahtjev je negativan, inače je neutralan. Što je veći s znači da je zahtjev pozitivniji, a što je manji zahtjev je negativniji.

Sentimentalnost– zanima nas naglašavanje emocija u zahtjevima. Dok su neki zahtjevi izrazito pozitivni i govore o veselim događajima drugi su izrazito negativni te naglašavaju životne negativnosti. Da bi odredili sentiment koristili smo paket vaderSentiment u Pythonu koji je razvijen u svrhu ocjenjivanja sentimenta na društvenim mrežama. Korištenjem preko 9 000 različitih lingvističkih svojstava koja uključuju emotikone, akronime i slang s naglaskom na sentiment ocijenjenih na skali -4 (izrazito negativno) - 4(izrazito pozivtivno) ocjenjuje se sentiment. Kao izlaz dobivamo ocjenu pozitivnosti, negativnosti, neutralnosti i ukupno izračunatu sentimentalnost zahtjeva. Mi smo koristili posljednju varijablu kao svojstvo sentimentalnosti.

Topic modeling - pročavanjem podataka ustanovili smo da neke priče naglašavaju nedostatak novca, neke emocionalne krize, a neke jednostavno glad. Uočili smo da bi oblikovanje tema moglo utjecati na naš ciljni rezultat. Da bi oblikovali glavne teme koristili smo NMF faktorizaciju (eng. Non-negative Matrix Factorization).

Nakon glavnog dijela preproceriranja slijedeći korak pripreme za NMF je bio tokeniziranje zahtjeva. Zahtjeve smo podijelili na individualne tokene koji predstavljaju izraze te smo pomoću njih konstruirali matricu zahtjeva. Time smo dobili nenegativnu matricu zahtjeva nad kojom ćemo vršiti dekompoziciju pomoću NMF faktorizacije. Izlaz su nam dvije matrice: jedna koja sadrži najfrekventnije riječi pojedine teme te druga koja sadrži opis teme. Testiranjem smo dobili da je najoptimalniji broj tema 7. Teme su zajedno s 10 najfrekventnijih riječi prikazane u tablici ispod.

TEMA	10 najfrekventnijih izraza
family	life, car, family, wife, bills, friends, mom, couple, girlfriend, husband
craving	pizza, craving, dinner, chesee, pepperoni, need, wanted, love, free, delicious
activity	comments, picture, photo, post, link, reason, game, graphic, video, reddit
money1	pay, paycheck, bank, check, cash, bills, money, broke, job, credit
appreciation	thanks, appreciate, favour, greatly, hoping, need, bless, kind, return, gift

money2	job, buy, work, bank, paid, rent, provide, card, bills, spent
student	hungry, broke, college, student, sob, studing, school, tuition, semestar, campus

3.2. Socijalni faktori

U našem skupu podataka postoji jako puno socijalnih atributa kao što su: status, duljina članstva, broj komentara, broj postova... Neki od njih su, logično, bitniji od drugih. Prilikom rješavanja zadatka jako je bitno odrediti u kojoj mjeri će oni utjecati na konačni rezultati. Na osnovu intuicije te raznim testiranjima odlučile smo se za tri najbitnija socijalna faktora.

Prvi su karma bodovi koje Reddit broji za sve postove i komentare pojedinog korisnika. U podatcima smo izlolirali polje *requester_upvotes_minus_downvotes_at_request.* Zatim smo izolirali polja *requester_account_age_in_days_at_request* i *requester_number_of_posts_on_raop_at_request* koji respektivno daju informaciju koliko je dugo korisnik član Reddit-a i da li je prije stavljao zahtjeve na stranicu *Random acts of pizza*.

4. Metoda potpornih vektora

Metoda potpornih vektora (SVM) je algoritam učenja korištenjem linearnih metoda u jezgrom induciranom prostoru značajki, pri čemu se kontrolira greška generalizacije korištenjem statističke teorije učenja i primjenjuje teorija optimizacije za rješavanje konveksnog kvadratnog programa na čije rješavanje se svodi učenje metodom SVM. Važno je obilježje SVM metode da je odgovarajući problem optimizacije konveksan, što jamči da rješenje nema lokalnog minimuma, odnosno, nađeno rješenje je globalni minimum.

Obzirom da je naš problem binarne klasifikacije linearno separabilan odlučili smo se za metodu potpornih vektora.

Kod korištenja SVM-a odabir parametara utječe na točnost zato je bitno izabrati pogodne parametre za naš skup podataka. Parametar C omogućuje razmjenu između preciznosti klasifikacije primjera za učenje i složenosti modela. Manje vrijednosti parametra C će dozvoljavati više pogrešnih klasifikacija primjera za učenje, a većim vrijednostima tog parametra bit će inducirani klasifikatori veće složenosti. Parametar γ definira koliko daleko doseže utjecaj jednog primjera. Niske vrijednosti parametra γ znače 'blizu' a visoke 'daleko'. Da bi došli do najoptimalnijih parametara koristili smo metodu *grid_search* iz paketa sklearn. Također nam je kao kernel predložila rbf.

5. Rezultati

Mjera točnosti je:

$$Acc = \frac{\#točno\_klasificirani\_zahtjevi}{\#ukupan\_broj\_zahtjeva}$$

Nakon pronalaska optimalnih C i γ na 1000 slučajno odabranih rezultata dobili smo slijedeće vrijednosti:

$$C = 10.0$$

$$\gamma = 1.0$$

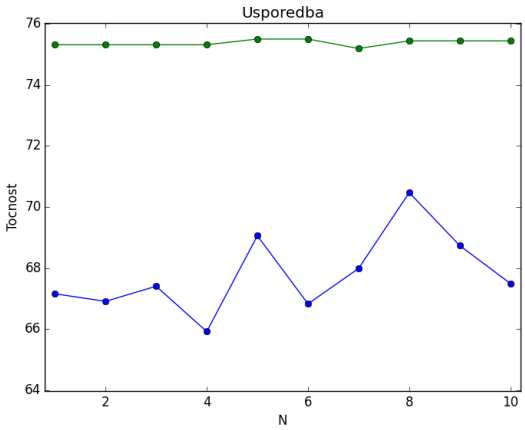
Koristeći 10-struku cross-validaciju dobili smo točnost 75.37%.

Ukoliko uzmemo druge parametre, na primjer

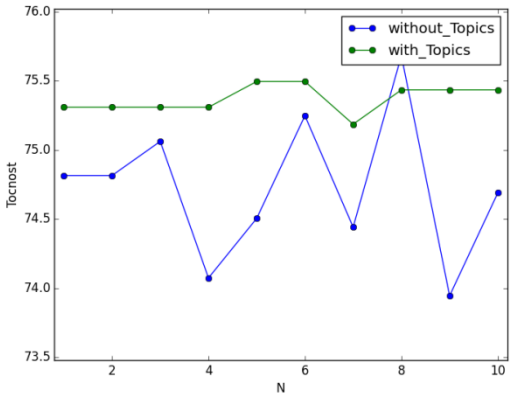
$$C = 100000.0$$

$$\gamma = 0.001$$

vidimo da je klasifikacija puno lošija.



Napomenuli smo da smo izabrali 7 glavnih tema. Testirali smo SVM sa različitim brojem i taj se pokazao najboljim. Također smo testiranjem dobili da 'topic modeling' ima utjecaj na sam rezultat. To pokazuje i slijedeći graf.



Također smo radili testiranje samo sa lingvistističkih i samo sa socijalnim faktorima. U oba slučaja rezultati su bili lošiji, no lingvistički faktori su se pokazali boljima u samostalnom radu.

Testiranjem smo dobili da recipročnost i evidentnost utječu na sam rezultat, dok sentimentalnost ima manji utjecaj. Kada smo izbacili sentimentalnost sama točnost se nije toliko smanjila koliko smo očekivali. Smatramo da je to posljedica toga što ljudi teško klasificiraju emocije u tekstu pa stoga je teško naučiti računalo na pravilnu klasifikaciju.

6. Zaključak

Cilj našeg projekta je bio predvidjeti čovjekovo djelovanje. U konačnici smo na osnovu lingvističkih i socijalnih faktora s točnošću 75.37% predvidjeli da li će neki korisnik dobiti besplatnu pizzu. Obzirom da smo koristili 10-struku cross-validaciju naš algoritam točno klasificira 301 od 404 zahtjeva. Smatramo da je to dobar rezultat jer ni sam čovjer ne može uvijek predvidjeti akcije drugog njemu nepoznatog čovjeka.

7. Literatura

[1] *How to Ask for a Favor: A Case Study on the Success of Altruistic Requests,* by Tim Althoff, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky

[2] Kaggle problem - *www.kaggle.com/c/random-acts-of-pizza*

[3] Strojno učenje - http://web.math.pmf.unizg.hr/nastava/su/