Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/22BE04Mijatovic

PREDIKCIJA OCENE APLIKACIJE NA OSNOVU KOMENTARA

PREDICTING APPLICATION RATINGS BASED ON USER REVIEWS

Marko Mijatović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – SOFTVERSKO INŽENJERSTVO I INFORMACIONE TEHNOLOGIJE

Kratak sadržaj – *U ovom radu predstavljena je specifikacija, implementacija i evaluacija sistema za predikciju ocene na osnovu tekstualnog komentara. Upoređena su dva pristupa – rekurentne neuronske mreže i modeli ansambla.*

Ključne reči: *mašinsko učenje, rekurentne neuronske mreže, modeli ansambla*

Abstract – *This paper presents a specification, implementation and evaluation of a system that predicts the rating of application based on textual comments. Two approaches were compared - recurrent neural networks and ensemble models.*

Keywords: *machine learning, recurrent neural networks, ensemble learning*

1. UVOD

Recenzije korisnika sadrže informacije koje su korisne za analitičare i dizajnere aplikacija, kao što su zahtevi korisnika, greške, i iskustva prilikom korišćenja aplikacija. Kao povratne informacije, recenzije korisnika govore prodavcima kako da poboljšaju kvalitet softverskog proizvoda. Potencijalni korisnici aplikacija najčešće pregledaju recenzije pre nego što ih instaliraju. Dakle loše recenzije, negativno utiču na percipirani kvalitet aplikacije, kao i na popularnost, a samim tim i zaradu.

Problem kojim se rad bavi definisan je na sledeći način: na ulazu se nalazi recenzija korisnika u tekstualnoj formi, dok izlaz treba da bude prediktovana ocena (broj 1-5).

Predikcija ocene na osnovu tekstualne recenzije može se posmatrati i kao regresioni, ali i kao klasifikacioni problem. Izučavajući literaturu, i radove koji se bave rešavanjem sličnog problema, razmatrani su različiti pristupi i upoređivani su na osnovu postignutih rezultata.

U radu "*Predicting star ratings based on annotated reviews of mobile apps.*", Monet i Dagmar su upotrebili sentiment analizu u kombinaciji sa modelima linearne regresije [1]. Ovaj rad je istakao koliko je sentiment analiza od velikog značaja za rešavanje problema. Iako je sentiment analiza vrlo moćna tehnika u obradi prirodnog jezika, njena primena na nepotpune skupove podataka vrlo verovatno će dati lošije rezultate.

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

Kao evaluaciona metrika za performanse primenjenih regresionih modela upotrebljena je RMSE (eng. Root Mean Square Error), što je vrlo dobar pokazatelj odstupanja prediktovanih vrednosti od pravih u problemu koji se rešava. Naime, drugi metodi evaluacije kao što su F1 mera i tačnost (eng. accuracy), koji se primenjuju kod klasifikacionih modela isto tretiraju i velike i male greške u predikciji modela. Nije isto da li je model prediktovao 2 za ocenu umesto 1, ili je prediktovao 5 umesto 1. Metrika evaluacije koja je upotrebljena u ovom radu, verodostojnije prikazuje odstupanje predikcija od stvarnih vrednosti, nego što to prikazuju metrike evaluacije korištene u drutim pristupima koji su razmatrani u nastavku.

S druge strane, Umer i dr. su u studiji "*Predicting numeric ratings for Google apps using text features and ensemble learning.*" koristili modele ansambla [2]. Pretprocesiranje podataka u ovom radu kvalitetno je odrađeno, na interesantan način (metod je predstavljen u samom radu) uklonjeni su autlajeri (eng. outliers), koji značajno mogu da naruše performanse primenjenih algoritama. Kao i u prethodnoj studiji upotrebljena je sentiment analiza kako bi se unapredile performanse pri predikciji. Modeli ansambla koji su korišteni su Random Forest, i GBM (eng. Gradient Boost Machine). Postignuta je tačnost 70 %, što se može smatrati vrlo dobrim rezultatom. Jedan od pristupa rešavanja problema koji će se koristiti u ovom radu su modeli ansambla.

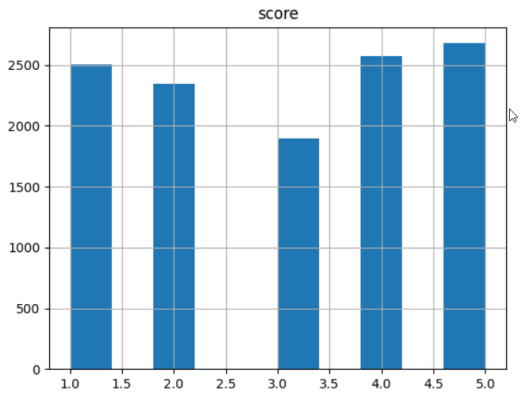
Gezici, Bahar i dr. su u radu "*Neural sentiment analysis of user reviews to predict user ratings.*" predstavili model rekurentne neuronske mreže koji je pokazao najbolje rezultate [3]. Postignuta je tačnost 87 %, što je značajno bolje u odnosu na rezultate pristupa gde su korišteni modeli ansambla. Najverovatnije je da je za ovako dobre rezultate zadužen ogroman skup podataka koji je korišten. Metod upotrebljen u ovom radu (rekurentne neuronske mreže) vrlo je pogodan za prirodu problema koji se rešava, te iz tog razloga postignuti su bolji rezultati nego u drugim studijama koje su razmatrane. Stoga, primarni pristup rešavanju problema koji je upotrebljen u ovom radu jesu rekurentne neuronske mreže.

Kako je skup podataka koji se koristi znanto manji nego skup korišten u gore navedenoj studiji, kvalitetno pretprocesiranje podataka je vrlo važno kako bi se postigli dobri rezultati.

2. METOD

Najpre je opisan skup podataka koji se koristio za treniranje i testiranje obučavanih modela mašinskog

učenja [4]. U skupu je 12496 recenzija. Ciljna varijabla je ocena, i posmatraće se balansiranost skupa podataka u odnosu na ciljnu varijablu. Na slici 1 prikazana je raspodela ciljne varijable – ocena korisnika.



Slika 1. *Histogram ocena*

Na osnovu slike 1, može se zaključiti da su podaci dobro balansirani. Raspodela ciljne varijable bi bila suprotnost u odnosu na Gausovu krivu. Prosečna ocena (3) je najmanje zastupljena, zatim imamo ocene (2) i (4), i najviše su zastupljene ocene (1) i (5). To donekle i ima smisla, jer najčešće recenzije ostavljaju ili oni korisnici koji su nezadovoljni aplikacijama, ili oni koji su oduševljeni.

2.1. Podela na trening/test skup

Skup podataka podeljen je na trening i test skup u razmeri 4:1. Za ovakvu raspodelu odlučeno je zbog toga što je skup podataka umerene veličine. Prilikom podele, očuvana je raspodela ciljne varijable u trening i test skupu. Za optimizovanje hiperparametara korištena je unakrsna validacija tako da nije bilo potrebe za posebnim izdvajanjem validacionog skupa.

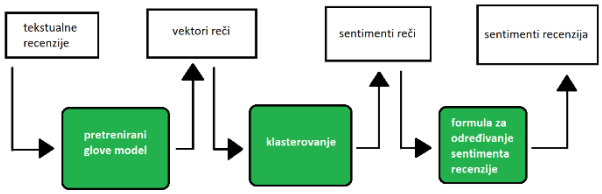
2.2. Uklanjanje autlajera

Recenzije koje se nalaze u korištenom skupu podataka većinski su napisane na engleskom jeziku, mada ima i onih koje su napisane na španskom, poljskom itd. Upotrebom *langdetect* biblioteke u programskom jeziku *Python*, pronađene su sve recenzije koje nisu napisane na engleskom jeziku i zatim su eliminisane. Autlajere je moguće ukloniti i vizualizacijom podataka, međutim u problemu koji se rešava tumačenje vizualizacije nema smisla. Autlajeri se moraju detektovati na neki drugi način. Ukoliko se sentiment recenzije ne poklapa sa ocenom koja je data (pozitivan sentiment i niska ocena, ili negativan sentiment i visoka ocena), onda takvu torku smatramo autlajerom. Dakle, potrebno je da se uoče sve recenzije kod kojih se sentiment ne poklapa sa ocenom. Nažalost, u korištenom skupu podataka ne postoji obeležje za sentiment recenzije, tako da treba uraditi predikciju sentimenta nenadgledanim učenjem. Sentiment analiza nenadgledanim učenjem predstavlja složen zadatak. Rafal Wojcik predstavio je svoje rešenje problema, koje je prilagođeno u ovom radu. U nastavku se opisuje ova metoda.

Prvi korak bio je da se reči iz recenzija prevedu u vektore realnih brojeva, na taj način da semantički slične reči imaju slične vrednosti vektora. Ova tehnika obrade prirodnih jezika poznata je kao *word embedding*. Dva dobro poznata *word embedding* algoritma su *word2vec* i *glove*. Kako je korpus reči svih recenzija u skupu podataka previše mali da bi se *word embedding* algoritam trenirao "od nule", upotrebljen je pretrenirani *Glove* model [10]. Model je izabran jer je bio treniran na podacima sličnim skupu podataka koji se koristio.

Drugi korak sentiment analize bio je da se klasteruju reči u dve grupe - sa pozitivnim i negativnim sentimentom. Svakoj reči pridružen je *sentiment score*, broj koji je označavao izraženost pozitivnog odnosno negativnog sentimenta. *Sentiment score* se računao tako što se delio broj klastera kome reč pripada (+1 za pozitivan, -1 za klaster sa negativnim sentimentom) sa distancom od centroida klastera. Na taj način formiran je rečnik sentimenata recenzija - svakoj reči iz recenzija pridružen je sentiment score.

Krajnji korak bio je da se odredi sentiment recenzije na osnovu sentiment rečnika dobijenog u prethodnom koraku. Svaka reč u recenziji pretvorena je u vektor brojeva upotrebom TF-IDF (eng. *term frequency-inverse document frequency*) algoritma, slika 2. Sentiment recenzije računao se kao skalarni proizvod dva vektora – *tfidf* težina reči u recenziji i vektora sentiment score-ova za svaku reč u recenziji. Znak sentimenta označavao je da li je u pitanju pozitivan ili negativan sentiment. Korisno je bilo što je za svaku recenziju dobijena numerička vrednost koja je predstavljala izraženost sentimenta.



Slika 2. *Koraci sentiment analize*

2.3. Pretprocesiranje teksta

Koraci u obradi teksta:

- uklonjeni su znakovi interpunkcije
- sva slova pretvorena su u mala (eng. *lowercase*)
- tokenizacija
- uklanjanje stop reči (eng. *stop words*)
- lematizacija umesto steminga, kako reči ne bi izgubile značenje

Neophodno je pretvoriti tekstualne recenzije u numeričku formu. *Word embedding* tehnike koje su korištene su TF-IDF i Glove. Mana tf-idf algoritma jeste što procesira jednu reč u trenutku, ne i okolinu reči tj. kontekst. Modeli ansambla trenirani su na recenzijama koje su pretprocesirane pomoću tf-idf algoritma, dok se kod rekurentne neuronske mreže u *embedding* sloju koristio pretrenirani Glove model.

2.4. Klasifikacija upotrebom modela ansambla

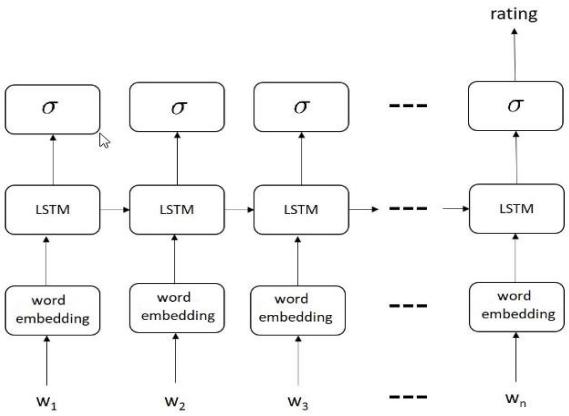
Problem rešavan u ovom radu može se posmatrati kao regresioni ali i klasifikacioni problem. U slučaju klasifikacije, u pitanju je multikategorijska klasifikacija. Primena binarnih klasifikatora moguća je i za probleme multikategorijske klasifikacije. Ideja ovog pristupa je da se obuči više modela, i da se njihovim glasanjem formira konačna predikcija.

Ansambl modeli na ulazu očekuju podatke u numeričkoj formi, pa je potrebno pretvoriti recenzije u numeričke podatke. U tu svrhu upotrebljen je TF-IDF algoritam, pripadnik *bag of words* grupe algoritama. Mane ovog algoritma su što se procecira jedna reč u datom trenutku (ne uzima se u obzir kontekst tj. okolina reči), a ne uzima u obzir ni semantičku analizu recenzija.

U poglavlju gde se opisuje uklanjanje autlajera, objašnjen je postupak sentiment analize nenadgledanim učenjem. Sentimenti dobijeni ovim postupkom uzeti su u obzir prilikom obučavanja ansambl modela. Upotrebljeni su bagging i boosting algoritmi: Random Forest, Ada Boost i Gradient boosting.

2.5. Rekurentna neuronska mreža

U kontekstu rešavanja problema u ovom radu, rekurentna neuronska mreža je idealan pristup, jer omogućuje procesiranje sekvenci. U nastavku je predstavljena struktura LSTM mreže koja je upotrebljena za rešavanje problema.



Slika 3. *Arhitektura mreže, preuzeto iz rada [3]*

Na slici 3 prikazana je arhitektura LSTM mreže. Na ulazu u model se očekuje sledeći format – reči predstavljene u vektorskom prostoru malih dimenzija koji definiše semantičke i sintaktičke osobine reči. Dakle, LSTM model na ulazu ima pretprocesirane reči recenzije, dok je izlaz sentiment recenzije. Za datu recenziju x, sa n reči $\{\omega_1, \omega_2, \dots, \omega_n\}$, najpre se svaka reč mapira na *word embedding* upotrebom pretreniranog Glove modela. Na izlaz svake LSTM jedinice primenjuje se sigmoid funkcija. Predikcija ocene je izlaz poslednje sekvence, pošto je ovakva arhitektura preslikavanje mnogo na jedan.

Rekurentna neuronska mreža trenirana je minimizacijom funkcije negativne verodostojnosti (eng. negative likelihood).

2.6. Hiperparametri modela

Za algoritam **Gradient boosting** hiperparametri su podešeni na sledeći način:

- *n_estimators* = 50,
  Predstavlja broj slabih klasifikatora, unakrsnom validacijom je utvrđena vrednost.
- *learning_rate* = 1,
  Brzina učenja, defaultna vrednost, nije promenjena jer se uobičajeno koristi.
- *max depth* = 15,
  Dubina stabla odlučivanja, hiperparametar koji kontroliše overfitting, jer veće dubine dozvoljavaju modelu da se previše prilagodi trening podacima, preporuka je da se parametar utvrdi unakrsnom validacijom, što je učinjeno.

Specifikacija hiperparametara za algoritam **Ada boost:**

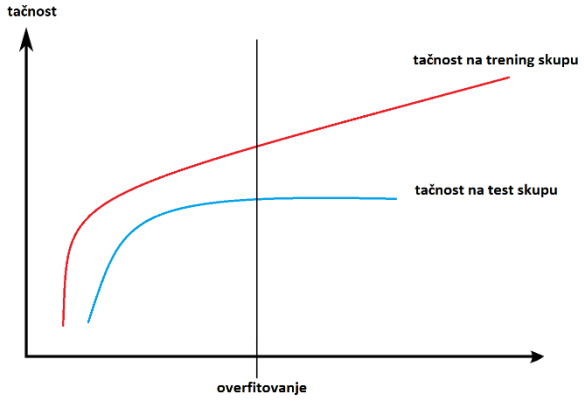
- *n_estimators* = 80,
  Predstavlja broj slabih klasifikatora, unakrsnom validacijom je utvrđena vrednost.
- *learning_rate* = 1,
  Brzina učenja, defaultna vrednost, nije promenjena jer se uobičajeno koristi.

Specifikacija hiperparametara za algoritam **Random Forest**:

- *n_estimators* = 100,
  Predstavlja broj slabih klasifikatora, unakrsnom validacijom je utvrđena vrednost.
- *min_samples_split* = 10,
  Minimalan broj uzoraka potreban da se razdvoji čvor stabla odlučivanja, određen unakrsnom validacijom.
- *min_samples_leaf* = 4,
  Minimalan broj uzoraka na svakom listu stabla odlučivanja, utvrđen unakrsnom validacijom.
- *max depth* = 30,
  Dubina stabla odlučivanja, hiperparametar koji kontroliše *overfitting*, jer veće dubine dozvoljavaju modelu da se previše prilagodi trening podacima, preporuka je da se parametar utvrdi unakrsnom validacijom, što je i urađeno.
- *bootstrap* = true,
  Hiperparametar koji govori kako će se popunjavati stabla odlučivanja, ako je postavljen na false, čitav skup podatka će se koristiti u svakom stablu odlučivanja, upotrebljena je defaultna vrednost, jer se uobičajeno koristi.

Specifikacija hiperparametara **rekurentne neuronske mreže**:

- *epochs* = 30,
  Predstavlja broj epoha treniranja neuronske mreže.
- *batch_size* = 4,
  Kontroliše koliko često se ažuriraju težine u neuronskoj mreži. Optimizuje se tako što se za fiksiran broj epoha isprobavaju različite vrednosti. Odabir ovog parametra zavisi dosta i od hardverskih kapaciteta mašine na kojoj se mreža obučava..



Slika 4. *Kriterijum za obustavljanje obučavanja mreže*

Rekurenta neuronska mreža trenirana je na računaru sa sledećim specifikacijama: procesor *Intel i7-10750H*, grafička kartica *NVIDIA GeForce GTX* 1650, 16 GB RAM memorije.

3. REZULTATI I DISKUSIJA

U tabeli 1, prikazani su rezultati primenjenog modela rekurentne neuronske mreže. Ovaj pristup dao je najbolje rezultate, što opravdava činjenica da ima *state-of-the-art* status u oblasti rešavanja problema.

Preciznost	Odziv	F1 mera	Tačnost
77.81	64.32	61.54	82.03

Tabela 1. *Performanse modela rekurentne neuronske mreže*

Algoritmi	Ada boost	Gradient boost	Random Forest
F1 мера	54.32	55.16	**57.88**
tačnost	67.78	68.89	**73.44**

Tabela 2. *Poređenje rezultata primenjenih modela ansambla*

Na osnovu tabele 2, vidi se da je od korištenih modela ansambla najbolje rezultate postigao algoritam Random Forest. Priroda problema koji se rešava je po svojoj složenosti pogodnija za *bagging* algoritme. Zbog toga su očekivani ovakvi rezultati, da je najbolje performanse postigao predstavnik *bagging* algoritama (Random Forest).

Dakle od svih upotrebljenih modela najbolje se pokazao model rekurentne neuronske mreže. Pretpostavlja se da bi rezultati bili još uspešniji, da se koristio veći skup podataka za obučavanje mreže. Ono što bi takođe moglo da unapredi performanse je *word embedding*, konkretno mogao bi se upotrebiti pretrenirani model koji je obučavan nad većim skupom podataka.

4. ZAKLJUČAK

U ovom radu predstavljen je sistem predikcije ocena na osnovu tekstualnih recenzija korisnika. Značaj za rešavanje ovog problema jeste u proceni kvaliteta softvera, koji je bitan kako prodavcima, tako i korisnicima softvera. Problem je rešen na dva različita načina, sa upoređivanjem performansi ova dva pristupa.

Prvi pristup bila je klasifikacija pomoću modela ansambla uz sentiment analizu. Drugi pristup rešavanju problema bio je upotrebom rekurentne neuronske mreže.

Upoređujući performanse, zaključeno je da se rekurentna neuronska mreža pokazala kao bolji pristup u rešavanju problema. Pokazano je koliko je *word embeding* važan za oba pristupa i u kojoj meri utiče na performanse.

Takođe sentiment analiza koja se oslanja na *word embeding*, značajno pomaže u predikciji ocene na osnovu tekstualnih komentara. Glavna prednost u rešenju predstavljenom u ovom radu jeste kvalitetno preprocesiranje podataka. Sve nepravilnosti i šumovi koji bi mogli da naruše performanse primenjenih algoritama su uklonjeni. U procesu otklanjanja outlajera, pojavio se novi složen problem – sentiment analiza nenadgledanim učenjem. Tom problemu moglo bi se posvetiti i u daljim istraživanjima jer ima veliki značaj i primenu.

Problem koji se rešavao mogao bi se i generalizovati. Pored aplikacija na Google Play prodavnici, ljudi recenziraju i ocenjuju mnoge druge stvrari. Obrada prirodnog jezika vrlo je značajna oblast, i automatizovana procena zadovoljstva korisnika na osnovu njegovog komentara je takođe od velikog značaja.

5. LITERATURA

[1] D. Monett and H. Stolte, “Predicting Star Ratings based on Annotated Reviews of Mobile Apps,” Oct. 2016, pp. 421–428. doi: 10.15439/2016F141.

[2] M. Umer, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, “Predicting numeric ratings for Google apps using text features and ensemble learning,” *ETRI Journal*, vol. 43, no. 1, pp. 95–108, Feb. 2021, doi: 10.4218/etrij.2019-0443.

[3] B. Gezici, N. Bolucu, A. Tarhan, and B. Can, “Neural Sentiment Analysis of User Reviews to Predict User Ratings,” in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, Samsun, Turkey, Sep. 2019, pp. 629–634. doi: 10.1109/UBMK.2019.8907234.

[4]https://www.kaggle.com/datasets/prakharrathi25/google-play-store-reviews

Kratka biografija:

Marko Mijatović rođen je 25. oktobra 1998. godine u Subotici. Završio je Osnovnu školu „Kizur Ištvan“ u Subotici 2013. godine odličnim uspehom, kao đak generacije. Gimnaziju „Svetozar Marković“ u Subotici završava 2017. godine. Diplomirao je na Fakultetu tehničkih nauka u Novom Sadu 2021. godine, na smeru Softversko inženjerstvo i informacione tehnologije.