

Sentiment analiza komentara na sajtu ženske odeće

Cvijetin Mladenović

Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad

mladenovic.cvijetin@gmail.com

Marko Radović

Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad

markoradovic996@gmail.com

Borislav Gajić

Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad

borawr22@gmail.com

Apstrakt— Proširenost interneta poslednjih decenija je omogućila generisanje velike količine podataka vezanih za prodaju ženske odeće. Klasifikacija komentara po njihovim ocenama je jedan od ključnih aspekata sistema za ocenjivanje oglasa za prodaju odeće. U ovom radu su istražene različite tehnike predikcija ocena oglasa na osnovu teksta komentara, kojima su korisnici davali pozitivne ili negativne utiske na objavama stranice. Korišćena su tri osnovna modela: *Convolutional Neural Network* (CNN), *Recurrent Neural Network* (RNN) i transformer. U kombinaciji sa CNN i RNN su isprobani kao reprezentativni (engl. *embedding*) slojevi pretrenirani *GloVe* vektori i *Word2Vec* model koji je treniran od nule nad trening skupom podataka. Pretrenirani transformer model je korišćen u okviru *feature-based* pristupa, odnosno kao reprezentativni sloj koji se ne trenira. Na izlaz transformera je postavljen klasifikator. Od svih isprobanih klasifikatora najbolje performanse je imao *Support Vector Machines* (SVM) sa linearnim kernelom. Kao metrika za evaluaciju performansi modela je korišćena F1 mera zbog nebalansiranosti ciljne labele. Konačno, na testnom skupu podataka RNN postiže 0,5298, CNN 0,5980 a transformer 0,6623 F1 meru.

Keywords— odeća, komentari, ocene, duboko učenje

I. UVOD

Baze podataka o ženskoj odeći mnogo su se obogatile u poslednjih par decenija, zahvaljujući širenju sajtova koji se bave prodajom odeće. Takvi sajtovi su danas veoma posećeni i korisnici ostavljaju veliki broj komentara i ocena za određene komade odeće. Time dobijamo mnogo informacija kroz komentare, njihove ocene i korisnička mišljenja što je veoma bitno za određivanje sentimenta komentara.

U ovom radu će biti predstavljeno par rešenja problema predikcije rejtinga komentara vezanih za žensku odeću, smatrajući pod rejtingom ocenu koju je tom komadu odeće dao neki od korisnika uz komentar. Za obučavanje modela koriste se komentari te se moraju primeniti razne tehnike obrade teksta pre samog obučavanja. Za sam model rešenje može obuhvatati par pristupa, poput metoda mašinskog učenja ili neuronskih mreža specijalizovanih za obučavanje nad tekstualnim podacima (*Natural Language Processing* tehnike).

Skup podataka je preuzet sa sajta *Kaggle* [2], sadrži 23486 komentara koji su korisnici ostavljali na različitim sajtovima za žensku odeću.

Konkretno metodologije koje su korišćene u radu su: Rekurentna i konvolutivna neuronska mreža sa pretreniranom *GloVe B6* reprezentacijom, kao i reprezentacijom koja je trenirana nad našim skupom podataka putem *Word2Vec* metoda. Pored navedenih korišćen je i transformer sa različitim klasifikatorima. Takođe je korišćen pristup generisanja dodatnih komentara unutar trening skupa putem *Pegasus* modela.

Najviše uspeha ostvareno je upotrebom transformera i *Support vector machine* (SVM) klasifikatora, dok smo nešto lošije rezultate dobili korišćenjem RNN i CNN mreža, kao i augmentacijom i korišćenjem dodatnih reprezentacija.

Detaljan opis skupa podataka, metoda pomoću kojih je problem rešavan i izazova na koje smo naišli prilikom obučavanja modela biće predstavljen u nastavku rada. U narednom poglavlju su ukratko izložena srodna istraživanja, odnosno radovi sa sličnom temom i metodikom. U trećem poglavlju biće predstavljen skup podataka koji je korišćen kao i način pripreme podataka za obučavanje i testiranje modela. Zatim se predstavlja metodologija koja se koristi za rešavanje problema klasifikacije komentara prema njihovim rejtingima. Nakon toga sledi prikaz rezultata sa različitim pristupima. Na kraju je predstavljen zaključak ovog rada.

II. SRODNA ISTRAŽIVANJA

A. Rad “*Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network*” [1] se bavi problemom određivanja da li dati komentar spada u pozitivan, negativan ili neutralan sentiment i da li se na osnovu zaključka iz komentara može ustanoviti da li se dati proizvod preporučuje za kupovinu ili ne. Autori su imali na raspolaganju 22621 javnih komentara [2].

Pristup koji je korišćen jeste bidirekciona rekurentna neuronska mreža sa *Long short-term memory* (LSTM). Iz komentara su uklonjeni graničnici poput “\n” i “\r”. Rejting koji je 3 ili više je računat kao pozitivna ocena dok je rejting manji od 3 uzet kao negativna ocena. Za mapiranje reči iz komentara na vektorski prostor, korišćena je *GloVe* reprezentacija.

Dokazano je da korišćenjem bidirekcionne mreže sa LSTM memorijom moguće dostići F1 meru od 0.88 za preporuku na osnovu komentara i 0.93 za klasifikaciju sentimenta. Upravo ova razlika u distribuciji klasa nastaje zbog nebalansiranosti pozitivnih i negativnih komentara, te klasifikacija izražava pristrasnost ka pozitivnim komentarima kojih ima više.

B. U radu “*Sentiment Analysis using machine learning algorithms: online women clothing reviews*” [3] autor koristi više metoda za rešavanje problema klasifikacije sentimenta. Korišćeni algoritmi u ovom radu su: *Support Vector Machine*, *Logistic Regression*, *Random Forest*, *Naive Bayes*. Model je evaluiran na osnovu tačnosti, preciznosti, odziva, F1 mere i *Area Under Curve* (AUC). Ukupna tačnost algoritama koji su korišćeni u ovom radu se može videti u tabeli 1.

Tabela 1. Prikaz dobijenih rezultata u radu “*Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network*”

Algorithm	Accuracy	Precision	Recall	F1	AUC
SVM	0,91	0,92	0,91	0,92	0,92
Logistic Regression	0,91	0,93	0,91	0,92	0,95
Random Forest	0,87	0,92	0,87	0,89	0,94
Naive Bayes	0,93	0,93	0,93	0,93	0,95

C. U trećem radu “*Fine-grained Sentiment Classification using BERT*” [4] proučavan je pristup višeklasne klasifikacije sentimenta naspram binarne/trinarne klasifikacije u prethodna dva rada.

Binarna klasifikacija može imati vrednosti 0 ili 1, odnosno pozitivne i negativne vrednosti, dok višeklasna može imati veći skup mogućih vrednosti, konkretno 5. Ovaj način sentiment analize se naziva *fine-grained*.

Za rešavanje problema je korišćen *Bidirectional Encoder Representations from Transformers* (BERT) model na koji je dodato nekoliko potpuno povezanih slojeva neuronske mreže. Eksperiment je izvršen nad *Stanford Sentiment Treebank fine-grained* (SST-5) skupom podataka. Kao metrika je korišćena tačnost. BERT model postiže tačnost od 55,5% a poređenja radi na ovom problemu bidirekcionni LSTM postiže tačnost od 49,1%.

D. Četvrti rad “*Self-Attention-Based BiLSTM Model for Short Text Fine-Grained Sentiment Classification*” [5] rešava problem trinarne klasifikacije, odnosno problem sentiment analize na tri klase, pozitivna, negativna i neutralna.

Evaluacija modela je izvršena nad tri skupa podataka sa *SemEval* 2014 Task 4 takmičenja. Korišćen model u radu je *Self-Attention based Bidirectional LSTM* (BiLSTM). Za inicijalizaciju težina vektora reči korišćen je pretrenirani *GloVe*, a model je isproban sa 50, 100, 200, 300 dimenzija vektora reči. Kao metrika se koristi tačnost.

Model predstavljen u radu postiže prosečnu tačnost od 73% na prvom skupu podataka, 65% na drugom i 65% na trećem skupu podataka. Model je upoređen sa običnim BiLSTM i običnim LSTM modelom koji postižu za nekoliko procenata niže rezultate.

III. OPIS SKUPA PODATAKA

U ovom poglavlju je opisan skup podataka [2], prikazani su atributi koji su dati za svaki komad odeće iz skupa podataka i prikazan deo eksplorativne analize koja je bila obavljena na prethodnom skupu podataka.

Skup podataka je preuzet sa sajta “Kaggle” i sadrži 23486 redova u kojima su opisani komentari. Sa pomenutog sajta je moguće preuzeti CSV fajl, koji sadrži sve podatke koji su korišteni. Za svaki komentar su dati određeni podaci, a to su:

- Clothing ID - jedinstveni identifikator objave komada ženske odeće
- Age - godine korisnika koji je ostavio komentar na objavi
- Title - naslov komentara
- Review Text - tekst komentara
- Rating - rejting koji je korisnik ostavio uz komentar gde je 1 najlošija ocena a 5 najbolja
- Recommended IND - binarna vrednost koja iskazuje da li korisnik preporučuje proizvod gde je vrednost 1 ili ne preporučuje gde je vrednost 0
- Positive feedback count - nenegativna vrednost koja označava broj korisnika koji su označili komentar kao pozitivan
- Division name - naziv nadodeljenja u koju se svrstava objava
- Department name - naziv odeljenja u koju se svrstava objava
- Class name - naziv klase u koju se svrstava objava

Iz skupa podataka je korišćen *Review Text* i *Rating*. Ostala polja su izostavljena jer nisu od interesa za rešavanje problema kojim se ovaj rad bavi.

Skup podataka je podeljen na tri podskupa, tako da postoje trening, test i validacioni skup podataka. Podela je odrađena na osnovu ciljne labele, atributa *Rating*, kako bi se očuvala ista distribucija ciljne labele u svakom od skupova podataka. Nakon podela trening skup se sadrži iz 22678 komentara, validacioni skup iz 2776 komentara a test skup se sadrži iz 2767 komentara. Podela je izvršena u razmeri 80/10/10.

Kako bi se poboljšale performanse modela, neophodno je očistiti tekst od nepotrebnih karaktera koji ne nose nikakvo značenje. U tu svrhu primenjeni su sledeći koraci:

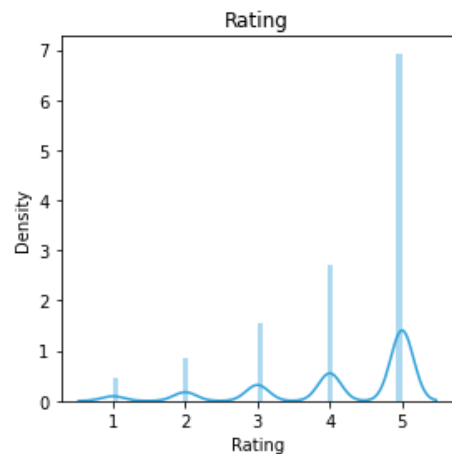
1. Uklanjanje praznih komentara
2. Izbacivanje duplikata komentara
3. Zamena (višestrukih) *whitespace* karaktera sa jednim razmakom
4. Izbacivanje potencijalnog HTML-a u komentarima upotrebom *BeautifulSoup* biblioteke
5. Zamena ili izbacivanje *unicode* karaktera
6. Izbacivanje nepoželjnih karaktera

Emoji nisu bili prisutni među komentarima, dok emotikoni jesu. Od svakog skupa podataka (trening, validacioni i testni) su napravljene tri verzije:

1. Sa emotikonima
2. Izbačeni emotikoni
3. Emotikoni zamenjeni sa kratkim tekstom koji najbolje opisuje njihovo značenje

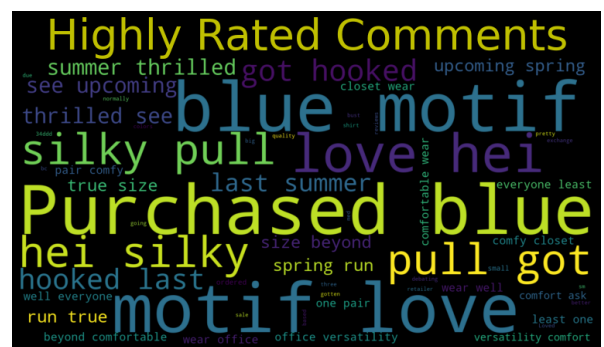
Ovo je urađeno kako bi se isprobalo koji pristup doprinosi poboljšanju performansi modela.

Eksplorativnom analizom je utvrđeno da u skupu podataka postoji značajno više komentara sa vrednošću rejtinga 5 u odnosu na komentare sa ostalim rejtinzima, što se može primetiti sa slici 1 na kojoj je prikazana raspodela komentara po rejtinzima, tačnije koliko komentara sadrži određeni rejting.



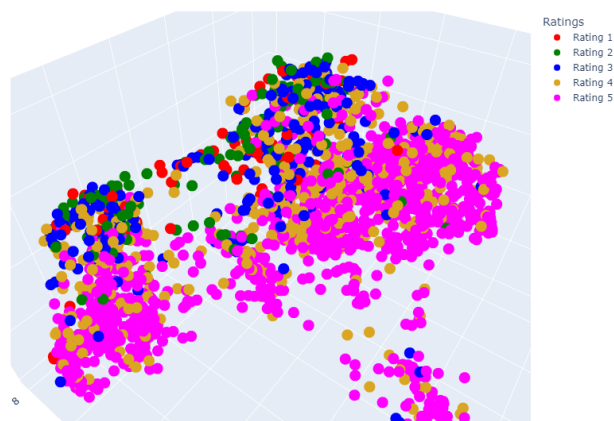
Slika 1. Prikaz zastupljenosti rejtinga komentara u čitavom skupu podataka

Utvrđeno je da su najčešće reči koje se pojavljuju “dress” i “petite”, a s obzirom da se u ovom radu bavimo klasifikacijom komentara na osnovu njihovog rejtinga, ovo nisu reči od velike važnosti te su one dodate u *stop words*. *Stop words* su reči koje ne nose posebno značenje u problemu koji se rešava ali pomažu pri razumevanju rečenice. Njihovim izbacivanjem se smanjuje broj bespotrebnih reči u komentarima što olakšava zadatak.



Slika 2. Prikaz wordcloud-a najučestalijih reči unutar trening skupa podataka

Na slici 3 su prikazani komentari u trodimenzionalnom prostoru. Komentari su vektorizovani korišćenjem pretreniranog transformera (detaljnije opisanog u metodologiji), potom su vektori spuštani na tri dimenzije upotrebom tehnike redukcije dimenzionalnosti algoritma *Uniform Manifold Approximation and Projection* (UMAP) [6]. Sa slike 3 se može videti da podaci iz različitih klasa nisu linearno separabilni u trodimenzionalnom prostoru, što pokazuje težinu rešavanog problema. Ipak, može se uočiti blagi trend prelaska sa viših ka nižim rejtinzima posmatrajući sliku sa desno na levo, što potvrđuje činjenicu da je model do neke mere uspeo da “razume” značenje teksta.



Slika 3. Prikaz vektora komentara u tri dimenzije

IV. METODOLOGIJA

Budući da su obučavajući podaci, odnosno komentari komada ženske odeće veoma dugački, autor skupa podataka je napravio ograničenje određenog broja karaktera da bi ulazni komentar bio koliko toliko približne veličine.

Nakon toga primenjuju se razne tehnike za obradu teksta. U prvobitnim pristupima se odmah primetilo da je mnogo veći broj komentara sa rejtingom 5 nego ostalih, s toga se pre svega morala primeniti augmentacija skupa podataka i to rečenica sa rejtingom manjim od 5. Posle augmentacije pristupljeno je raznim metodama i varijacijama za obučavanje modela i to kroz rekurentnu, konvolutivnu i transformer modele neuronskih mreža.

Da bi se podaci pripremili za ulaz u neuronske mreže koje su trenirane od nule, odrađena je lematizacija, gde je svaka reč unutar rečenica komentara zamenjena njenom osnovom reči. Na taj način se povećala mogućnost da će se iste reči pojavljivati u više komentara, jer reči koje imaju isti osnovni oblik imaju i isto značenje. Time će se smanjiti vokabular i omogućiti više reči različitog značenja da se nađu u vokabularu. Za lematizaciju je korišćena biblioteka *spacy* [7].

Nakon što su podaci preprocesirani, spremni su za algoritme mašinskog učenja. Algoritmi i metode koji su korišćeni za rešavanje problema izloženog u radu su:

- A. Transformer
- B. Rekurentna neuronska mreža
- C. Konvolutivna neuronska mreža
- D. Pretrenirani *GloVe embedding* uz CNN
- E. Trenirani *Word2Vec embedding* uz CNN

A. Nakon rada u kome je objavljen BERT [8] model nastaje procvat u oblasti obrade prirodnog jezika. Ovaj model postavlja nove rekorde u mnogim specifičnim NLP oblastima. Da bi se ispitalo da li će transformer model imati bolje performanse od ostalih korišćenih modela na problemu koji se rešava u ovom radu korišćen je *nli-mpnet-base-v2* transformer model iz *sentence-transformer* [9] biblioteke. Osnovu ovog modela čini *Masked and permuted pre-training* (MPNET) [10] model koji je slične arhitekture kao BERT, ali prevazilazi neke od mana BERT modela.

Pretrenirani transformer model je korišćen kao sloj za vektorizaciju teksta koji se ne obučava, dok se na izlaz ovog modela dodaje klasifikator koji se trenira i daje konačnu predikciju rejtinga za dati komentar. Isprobani klasifikatori su tranirani nad trening skupom a evaluirani nad validacionim skupom podataka sa emotikonima. Skup podataka sa emotikonima je izabran iz razloga što je MPNET model pretreniran nad velikom količinom podataka sa interneta koji su najverovatnije sadržali emotikone, pa je moguće da je model naučio da ih razume.

Za svaki klasifikacioni model su pronađeni približno najbolji hiperparametri. Klasifikacioni modeli su preuzeti iz *scikit-learn* [11] biblioteke. Klasifikacioni model koji je postigao najbolje rezultate na validacionom skupu će biti iskorićen za krajnju evaluaciju nad testnim skupom.

Kako bi se performanse najboljeg modela potencijalno još dodatno poboljšale, primenjena je augmentacija trening skupa podataka. Kao model za augmentaciju je korišćen *Pegasus* [12]. To je model koji se koristi za problem apstraktne sumarizacije.

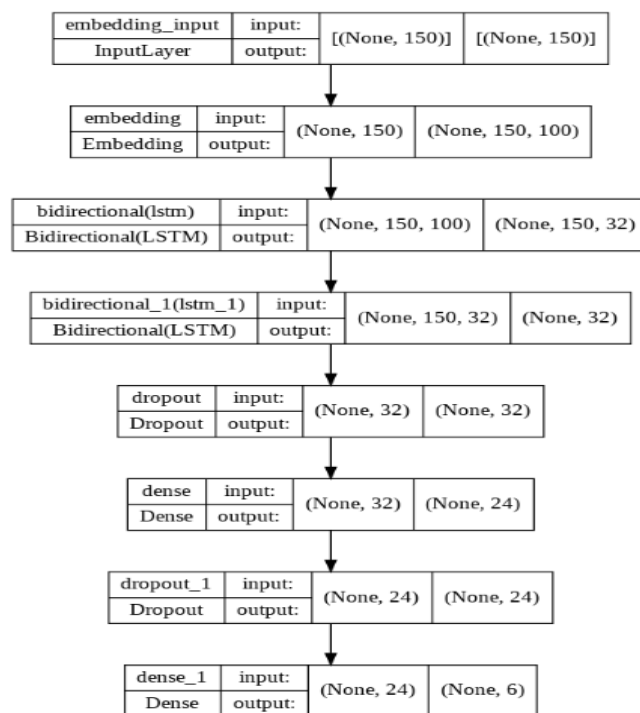
Isprobani pristupi:

1. Za svaki komentar iz trening skupa se generiše po jedna parafraza (da bi se ravnomerno povećao obučavajući skup)
2. Za svaki komentar koji nema rejting 5 se generiše po 5 parafraza (da bi izbalansirali broj komentara sa nižim rejtingom)

B. Rekurentne neuronske mreže ne uče isključivo reprezentaciju za svaku reč, nego i reprezentaciju cele rečenice. Korisne su jer čuvaju informaciju o trenutnom obeležju, kao i susedna obeležja za predviđanje, ali nekad ne uspeju da uhvate susednu reprezentaciju drugih sekvenci. LSTM može pomoći u rešavanju ovog problema jer razume kontekst zajedno sa nedavnom zavisnošću. Otuda su LSTM posebna vrsta rekurentnih neuronskih mreža gde razumevanje konteksta može biti korisno.

Korišćena su dva sloja LSTM, oba sloja su bidirekcionni LSTM sa po 16 neurona. Njima je praćen *Dropout* sloj od 0.2 kako bismo sprečili pretreniranost (engl. *Overfitting*). Bidirekcionni LSTM kao što mu i samo ime govori je dvosmerna mreža, odnosno ima pristup i prošlim i budućim ulaznim funkcijama za određeno vreme. Ovo je posebno važno pri označavanju sekvenci, stoga može poboljšati performanse modela na problemima klasifikacije sekvenci.

U problemima gde su dostignuti svi vremenski koraci ulazne sekvence, dvosmerni LSTM obučava dva umesto jednog LSTM na ulaznoj sekvenci. Pridodat je i *Dense* sloj od 24 neurona koji koristi *Rectified Linear Unit* (ReLU) aktivacionu funkciju, praćen sa još jednim *Dropout* slojem od 0.2 uz završni *Dense* sloj sa brojem izlaznih neurona jednak broju jedinstvenih rejtinga i *softmax* aktivacionom funkcijom sa korakom učenja od 0.001.



Slika 4. Arhitektura rekurentne neuronske mreže

C. Konvolutivne neuronske mreže se baziraju na klasifikaciji ulaznih podataka, tako da izlaz može biti jedna ili više klasa. Ideja konvolutivnih mreža je da se postavi veći broj slojeva za otkrivanje bitnih karakteristika ulaznih podataka.

Ulaz u mrežu je sekvenca 256-dimenzionalnih vektora, gde svaki reprezentuje reč iz komentara. Mreža uči vektor reprezentacije za vokabular reči koji uključuje čitav skup tokenizovanih reči.

Kod konvolutivne mreže imamo konvolutivni sloj, sloj sažimanja i potpuno povezane slojeve. Veze između ovih slojeva su aktivacione funkcije, u ovom slučaju to su *ReLU*, *LeakyReLU* i *Softmax*. *LeakyReLU* je bio od posebnog značaja iz razloga što se negativna vrednost množi sa izabranim koeficijentom $\alpha=0.1$ i time su se sačuvali moguće dobre karakteristike za naredni sloj sažimanja. Sažimanje maksimumom (engl. *MaxPooling*) je važan koncept konvolutivnih neuronskih mreža jer se uklanjanjem vrednosti, koje nisu maksimalne, smanjuje izračunavanje. *Pooling* sloj se obično dodaje nakon konvolucionog sloja uglavnom radi smanjenja dimenzionalnosti mape karakteristika radi računarske efikasnosti, što kasnije može poboljšati stvarne performanse.

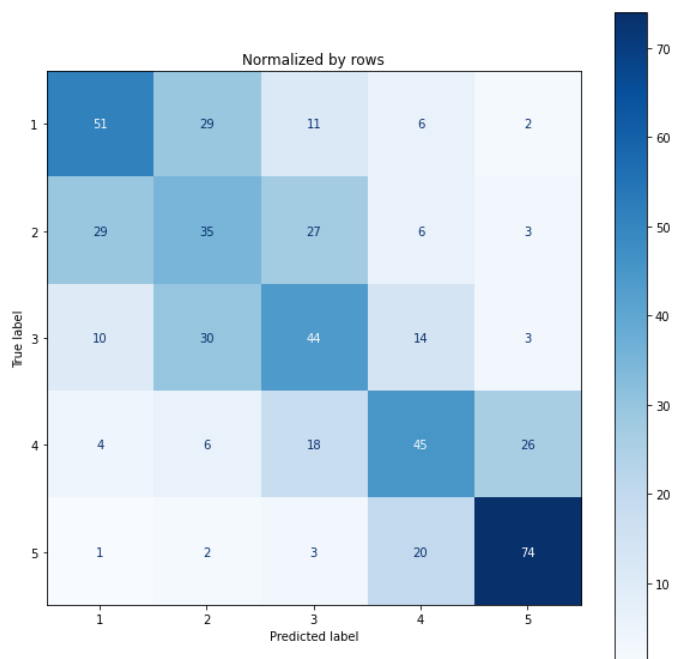
V. Analiza grešaka

U cilju boljeg razumevanja predikcija modela samim tim i potencijalnog unapređenja performansi modela, primenjena je analiza grešaka. Analiza grešaka je izvršena nad predikcijama transformer modela za validacioni skup sa emotikonima, koji je korišćen za izbor klasifikacionog modela. Prvi korak je analiza klasifikacionog izveštaja koji je prikazan u tabeli 3.

Tabela 2. Klasifikacioni izveštaj

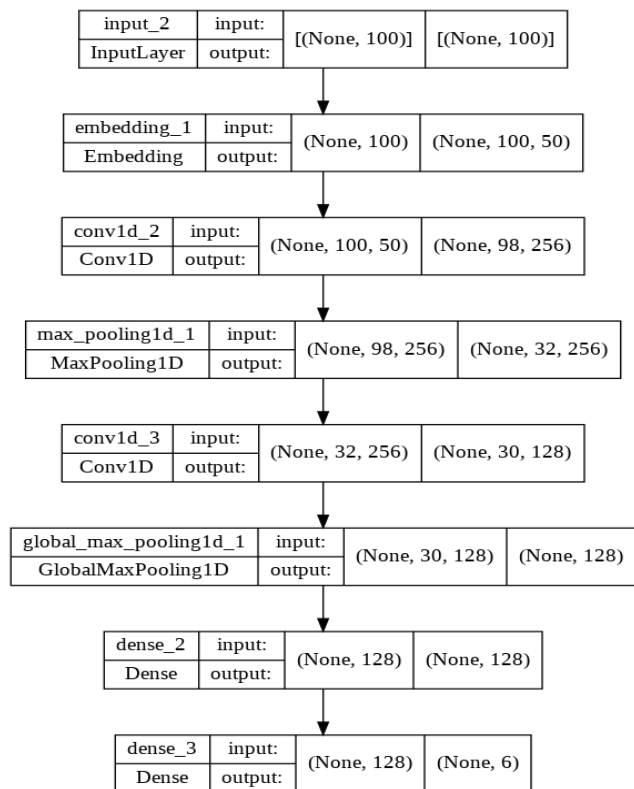
gold label	precision	recall	f1-score	support
1	0.28	0.51	0.36	82
2	0.26	0.35	0.30	157
3	0.40	0.44	0.42	282
4	0.43	0.45	0.44	495
5	0.86	0.74	0.80	1246

Iz table se vidi da model najbolje radi za komentare čiji je rejting 5, dok za sve ostale komentare radi dosta lošije. Razlog ovome najverovatnije leži u tome što najviše ima komentara sa rejtingom 5 pa je model imao najviše primeraka iz te klase za obučavanje. Da bi se uočio odnos između prediktovane i ciljne labele, odnosno, koje labele model najčešće prediktuje za koje ciljne labele, prikazana je matrica konfuzije na slici 6.



Slika 6. Matrica konfuzije u procentima, normalizovana po ciljnim labelama

Posmatrajući matricu konfuzije može se zaključiti da model, kada greši najčešće prediktuje za jedan jači ili slabiji rejting. Moguć razlog za to je što nije lako odrediti tačan



Slika 5. Arhitektura konvolutivne neuronske mreže

D. Pojava vektorskih reprezentacija reči, kao što su *GloVe* i *Word2Vec*, dovela je do vidljivih napredaka u raznim sferama procesiranja govornih jezika. Kod vektorskih reprezentacija, reči sa sličnim kontekstima imaju slične reprezentacije, tj. slične reči su blizu u vektorskom prostoru, a različite daleko.

Kontekst reči predstavljaju reči koje okružuju datu reč. Prednost *GloVe*-a u odnosu na *Word2Vec* je što se *GloVe* ne oslanja samo na lokalne informacije o kontekstu reči već gleda globalnu statistiku pojavljivanja reči za dobijanje vektora reči.

Korišćene su pretrenirane *GloVe 6B* reprezentacije vektora za embedding koju pruža univerzitet *Stanford*.

E. Za razliku od *GloVe* prezentacije, *Word2Vec* se oslanja samo na lokalne informacije o jeziku. To jest, na semantiku naučenu za datu reč utiču samo okolne reči. Koristili smo *Word2Vec* model iz biblioteke *gensim.models*. Model prima preprocesirane reči, obrađuje i obavlja vektorizaciju u paraleli radi postizanja brzih performansi. Nakon obučavanja modela iz *gensim.models* biblioteke potrebno je formirati reprezentacioni vektor i proslediti dalje u našu konvolutivnu neuronsku mrežu. Nakon 5 uzastopnih treniranja modela, rezultati su u proseku 4.5% lošiji u odnosu na rezultate dobijene korišćenjem embeddinga kod obične CNN mreže.

rejting za komentar koji nije ni potpuno pozitivan ni potpuno negativan.

Da bi se bolje razumelo zašto model greši u nastavku će biti prikazani primeri komentara iz validacionog skupa podataka.

Ciljna labela: 1, prediktovana labela: 5

"Got this in purple and love it! ordered red and silver and only received red so far, but had to return because it was oddly shaped and didn't fit well like the purple. not sure what happens. maybe miss sized?"

Ciljna labela: 2, prediktovana labela: 5

"I loved this dress online and was luckily able to locate it in a store about an hour away from me. i drove to the store to purchase it; however, when i tried it on, i was so disappointed! the print is adorable and it is made of great quality, but the swing style is definitely more than mini! i could have fit 3 of myself in this dress. it is definitely not as depicted in the photographs."

U ovim situacijama kada model značajno greši, može se primetiti da komentar počinje sa jednim sentimentom a završava sa drugim. Model najčešće prediktuje početni sentiment.

Ciljna labela: 5, prediktovana labela: 4

"Super cute, well made, great embroidery. waist runs big."

"I like this dress a lot. it does not have much form but with a belt looks great!"

Skoro potpuno pozitivan sentiment dok postoji jedna mala zamerka, koja najverovatnije navede model da prediktuje manji rejting od 5.

Ciljna labela: 4, prediktovana labela: 5

"I love this vest, good denim quality"

"Tried this on in the store, very soft and hangs well, better on, than the hanger. needless to say i bought it!"

Posmatrajući tekst komentara, nije jasno zašto bi korisnik dao ocenu koja je manja od 5. Ovo su verovatno primeri outlier-a.

Ciljna labela: 2, prediktovana labela: 1

"It was huge. material looked good but was too bulky. will be returning it."

"Cute in the on line photo but in person it looks cheap and has gold running through it. if you look really close in the photo you can see it."

U ovim primerima dominira negativan sentiment dok se na jednom mestu u komentaru kupac ponadao da će biti zadovoljan kupovinom. Verovatno je to razlog zašto ciljna labela nije 1.

Ciljna labela 2, prediktovana labela: 3

"I loved the look of this sweater on-line, but it just didn't work for me. the fit wasn't good enough. the style okay. the color fine. the feel, scratchy. overall, it did not work for my body."

"This dress looks very beautiful on the model shown. unfortunately, when i tried it on, it did not flow as nicely as i would have liked it to. i also didn't care much for the waist ties on the side. an elastic waist might have been better."

Komentari ovog tipa najčešće počinju sa pozitivnim sentimentom a završavaju sa negativnim. Karakteristično je što komentar sadrži dosta pozitivnih reči, što je moguć razlog zašto model prediktuje veći rejting od ciljnog.

Na osnovu gore navedenih primera, može se zaključiti da je ovaj problem težak za rešavanje čak i za čoveka. Nije lako odrediti tačnu granicu između određenih rejtinga. Takođe, u dosta primera nije moguće odrediti tačan rejting samo na osnovu komentara koji je kupac ostavio, već je u konačnom rejtingu verovatno učestvovao i subjektivni osećaj kupca koji nije u potpunosti prenesen na tekst komentara.

Pokušaj rešavanja ovog problema je augmentacija trening skupa, tj. dodavanje primera u trening skup, kako bi model bolje naučio razlike između rejtinga.

VI. Rezultati

Svi modeli masinskog učenja su trenirani nad istim trening skupom i testirani nad istim test skupom podataka. Način na koji su podaci dobavljeni i procesirani je pomenut u poglavlju III. Kao metrika za upoređivanje korišćena je F1 mera (engl. *F1 measure*). Da bi mogli da objasnimo F1 meru, potrebno je prethodno objasniti preciznost (engl. *Precision*), tačnost (engl. *Accuracy*) kao i odziv (engl. *Recall*). Preciznost predstavlja udeo dobro predviđenih primera određene klase u ukupnom broju primera koje je model svrstao u datu klasu. Time dobijamo meru koji deo rezultata u jednoj klasi je uspešno klasifikovan. Odziv predstavlja osetljivost modela, odnosno, pokazuje koliko je relevantnih rezultata algoritam vratio.

Odziv se računa po formuli:

$$Recall = \frac{True\ positives}{True\ Positives + False\ Negatives}$$

Tačnost se računa po formuli:

$$Accuracy = \frac{True\ positives + True\ Negatives}{All\ Samples}$$

Preciznost se računa po formuli:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

F1 mera kombinuje preciznost i odziv i računa se po sledećoj formuli:

$$F1\ Score = \frac{2 \times (Precision \times Recall)}{Pre}$$

Kako bi se izabrao klasifikator koji daje najbolje rezultate u kombinaciji sa transformer modelom, evaluirani su različiti klasifikatori nad validacionim skupom podataka sa emotikonima. Rezultati su prikazani u tabeli 2.

Tabela 3. Rezultati za transformer model u kombinaciji sa različitim klasifikacionim modelom.

Klasifikator	Hiperparametri	micro F1
SVM sa linearnim kernelom	random_state=1	0.6732
SVM sa linearnim kernelom	class_weight='balanced'	0.6065
SVM sa rbf kernelom		0.6653
K Neighbors Classifier		0.5897
Random Forest Classifier	n_estimators=300, max_depth=10	0.6153
MLP Classifier	alpha=0.01, early_stopping=True	0.6697
Logistic Regression CV	multi_class='multinomial'	0.6644
Linear Regression		0.6091

Iz tabele se može videti da najbolji rezultat postiže SVM sa linearnim kernelom. Slične performanse postižu: SVM sa rbf kernelom, MLP Classifier i Logistic Regression CV. Lošije rezultate postižu: SVM sa linearnim kernelom kome je class_weight='balanced', K Neighbors Classifier, Random Forest Classifier i Linear Regression.

Da bi se poboljšale performanse modela, korišćena je augmentacija. Dva korišćena pristupa su detaljnije opisana u poglavlju metodologije. Prvi pristup minimalno kviri performanse modela dok drugi pristup značajno spušta performanse modela. Potencijalan razlog za to je što augmentacijom nisu dodati potpuno novi primeri u obučavajući skup, već semantički vrlo slični postojećim podacima.

Rezultati za sve modele korišćene u ovom radu su prikazani u tabeli 4 za čiju validaciju je korišćena F1 mera.

Kako bi rezultati bili što merodavniji, uzet je prosek od 5 obučavanja nad istom arhitekturom i parametrima za svaki od modela koji su korišćeni za rešavanje problema predikcije rejtinga po sledećoj formuli:

$$\text{Result} = \frac{R1 + R2 + R3 + R4 + R5 + \dots RN}{N}$$

Rekurentna neuronska mreža se za ovaj problem nije pokazala kao najbolje rešenje. Ako pogledamo tabelu možemo uočiti da rezultati za F1 meru imaju u proseku tačnost od 0.5298. Radi sprečavanja pretreniranosti mreže iskorišćen je *Early stopping*. Još jedna od mana jeste ta što je rekurentna mreža osetljivija na činjenicu da preovlađuje veći broj komentara sa rejtingom 5, stoga u većini slučajeva kao izlaz predviđa rejting 5.

Konvolutivna mreža pokazuje bolje rezultate od rekurentne. Njen prosek je 0.5980. Smanjenjem broja parametara može se kontrolisati i pretreniranost (engl. *Overfitting*) mreže. Radi boljih rezultata u odnosu na rekurentnu mrežu, pokušana je upravo nad njom augmentacija trening skupa kao i pretrenirani *GloVe* i *Word2Vec* pristup koji smo trenirali nad našim trening skupom podataka. Dok se augmentacija i *Word2Vec* pristup nisu pokazali kao bolja rešenja, pretrenirana *GloVe* reprezentacija je dala skoro identične rezultate u odnosu na bazičnu konvolutivnu mrežu. Moguć razlog za slične performanse se nalazi u činjenici da veći skup reči sa sobom donosi nova obeležja iz oblasti odeće koja se nisu nalazila u početnom skupu.

Najbolje rezultate nad testnim skupom postiže transformer model. Za krajnju evaluaciju u kombinaciji sa *nli-mpnet-base-v2* transformer modelom je izabran SVM sa linearnim kernelom kao klasifikator koji se najbolje pokazao nad validacionim skupom. Rezultati pokazuju da je za ovaj problem korisnije koristiti transfer learning nego trenirati mrežu od nule. Moguće je da je transformer model u svojim težinama sačuvao mnogo više informacija nego što su to mogle da nauče RNN i CNN mreže tokom treniranja. Druga mogućnost zašto transformer postiže značajno bolje rezultate od ostalih modela je to što je on dosta kompleksniji model i bolje "razume" semantiku teksta što su pokazali dosadašnji rezultati u drugim radovima.

Tabela 4. Rezultati nad testnim skupom podataka sa uklonjenim emotikonima

Model	F1 mera
RNN	0,5298
CNN	0,5980
CNN + <i>GloVe pretrained embedding</i>	0,5979
CNN + <i>Augmented dataset</i>	0,5121
CNN + <i>Word2Vec embedding layer</i>	0,5592
Transformer + SVM	0.6623

Kako bi se ispitalo da li prisustvo emotikona u komentarima olakšava rešavanje problema predviđanja rejtinga komentara, transformer model je treniran i evaluiran nad tri verzije skupa

podataka: sa emotikonima, bez emotikona i zamenjenim emotikonima sa kratkim tekstom koji ih najbolje opisuje. Rezultati su prikazani u tabeli 5.

Tabela 5. Rezultati za transformer model za tri različita skupa podataka

Skup podataka	F1 mera
Sa emotikonima	0.6628
Bez emotikona	0.6623
Zamenjeni emotikoni sa kratkim tekstom	0.6619

Na osnovu rezultata može se zaključiti da prisustvo emotikonima u komentarima minimalno utiče na poboljšanje performansi modela. Pretpostavka je da je model tokom *pretraining* faze bio izložen tekstu sa prisustvom emotikona, te je delimično naučio da “razume” njihovo značenje.

VII. Zaključak

U ovom radu su prikazane različite tehnike rešavanja problema predikcije rejtinga komentara vezanih za žensku odeću. Skup podataka se sastojao od velikog broja komentara. Korišćeni atributi su tekst komentara i rejting. Na početku je izvršena eksplorativna analiza, izbačene su stop reči, izvršena je lematizacija. Posle preprocesinga skup podataka je bio spreman da se podvrgne određenom skupu algoritama, koji su vršili predikciju rejtinga na osnovu teksta komentara. Pre ovoga je skup podataka podeljen na obučavajuće podatke, validacione, kao i test podatke.

Za predikciju rejtinga korišćeno je nekoliko algoritama. Za upoređivanje je korišćena F1 mera. Korišćenjem ove mere može se zaključiti kakve rezultate daju pojedini algoritmi na skupu podataka čije su labele nebalansirane. Lošiji rezultati su dobijeni korišćenjem rekturentne i konvolutivne neurosnke mreže. Nešto bolji rezultati dobijeni su upotrebom transformera u kombinaciji sa SVM što se moglo očekivati jer je transformer značajno kompleksniji model od prethodna dva.

Kao potencijalna poboljšanja upotrebljenih algoritama, upotrebljena je pretrenirana *GloVe* reprezentacija kao ulazni vektor u CNN i RNN. Ovaj pristup je dao neznatno skoro identične rezultate u odnosu na standardnu reprezentaciju koja je uspostavljena nad našim skupom podataka. *Word2Vec* reprezentacija nad našim skupom podataka je dala primetno lošije rezultate. Augmentacija trening skupa podataka za rejtinge od 1-4 takođe nije dala bolje rezultate.

Analizom grešaka utvrđeno je da problem sam po sebi nije lako rešiv ni za čoveka. Komentari neretko započinu pozitivno te na kraju završe u negativnoj konotaciji, ili pak obratno. Takođe postoje i oni komentari koji po svojoj sadržini ne odgovaraju dodeljenom rejtingu. Primer za to je komentar koji je u potpunosti pozitivan a dodeljeni rejting nije 5. Dosta negacija u komentarima je predstavljao dodatan problem za jednostavnije modele kao što su CNN i RNN.

Dalje unapređenje performansi bi se moglo postići prilagođavanjem (*fine-tuning*) transformer modela trening skupu podataka. Ovaj pristup se u praksi do sada pokazao kao jedan od najboljih za rešavanje problema klasifikacije teksta. Takođe, što se tiče RNN i CNN bilo bi veoma zanimljivo napraviti matricu reprezentacija sa većim težinama dodeljenim manjim vrednostima rejtinga jer su modeli preferirali ocenjivanje komentara visokim ocenama zbog nebalansiranosti skupa podataka, čak i nakon augmentacije podataka tj. povećanja broja komentara sa niskim vrednostima rejtinga.

LITERATURA

- [1] Agarap, A. F. (2018). Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network (rnn). arXiv preprint arXiv:1805.03687.
- [2] Women's E-Commerce Clothing Reviews skup podataka <https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews>
- [3] Xie, S. (2019). Sentiment Analysis using machine learning algorithms: online women clothing reviews (Doctoral dissertation, Dublin, National College of Ireland).
- [4] Munikar, M., Shakya, S., & Shrestha, A. (2019, November). Fine-grained sentiment classification using bert. In 2019 Artificial Intelligence for Transforming Business and Society (AITB) (Vol. 1, pp. 1-5). IEEE.
- [5] Xie, J., Chen, B., Gu, X., Liang, F., & Xu, X. (2019). Self-attention-based BiLSTM model for short text fine-grained sentiment classification. IEEE Access, 7, 180558-180570.
- [6] McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
- [7] Spacy biblioteka <https://spacy.io/>
- [8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- [9] Sentence-transformers repozitorijum <https://github.com/UKPLab/sentence-transformers>
- [10] Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnnet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems, 33, 16857-16867.
- [11] Scikit-learn biblioteka <https://scikit-learn.org/stable/index.html>
- [12] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning (pp. 11328-11339). PMLR.