Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 621.9
DOI: https://doi.org/10.24867/29BE16Mladjenovic

NAPREDNE TEHNIKE ZA SENTIMENT ANALIZU: STUDIJA KLASIFIKACIONIH I GENERATIVNIH MODELA NA ONLINE KOMENTARIMA

ADVANCED TECHNIQUES FOR SENTIMENT ANALYSIS: A STUDY OF CLASSIFICATION AND GENERATIVE MODELS ON ONLINE COMMENTS

Cvijetin Mlađenović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Ovaj rad se bavi sentiment analizom komentara korisnika koristeći različite tehnike obrade prirodnog jezika i algoritme dubokog učenja. Implementirani su modeli poput rekurentnih neuronskih mreža (RNN), konvolutivnih neuronskih mreža (CNN), Word2Vec, GloVe, BERT, kao i generativni llama-7b-hf model. Skup podataka je preuzet sa Kaggle platforme i sadrži komentare korisnika na različite proizvode. Evaluacija modela je izvršena korišćenjem F1 mere, omogućavajući detaljnu analizu performansi u kontekstu nebalansiranih klasa. Najbolji rezultati su postignuti upotrebom transformera i SVM klasifikatora*.

Ključne reči: *Sentiment analiza, RNN, CNN, Word2Vec, GloVe, BERT, llama-7b-hf, NLP, klasifikacija*

Abstract – *This paper explores sentiment analysis of user comments using various natural language processing techniques and deep learning algorithms. Implemented models include Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Word2Vec, GloVe, BERT, and the generative llama-7b-hf model. The dataset was obtained from the Kaggle platform and contains user comments on various products. Model evaluation was performed using the F1 score, enabling a detailed analysis of performance in the context of imbalanced classes. The best results were achieved using transformers and SVM classifiers.*.

Keywords: *Sentiment analysis, RNN, CNN, Word2Vec, GloVe, BERT, llama-7b-hf, NLP, classification*

1. UVOD

Razvoj interneta i digitalnih tehnologija značajno je promenio način na koji ljudi komuniciraju i kupuju. Sentiment analiza ima široku primenu, od poslovne inteligencije i marketinških istraživanja do unapređenja korisničkog iskustva. U *e-commerce* industriji, ona omogućava trgovcima da prate zadovoljstvo kupaca, identifikuju probleme i unaprede svoje proizvode i usluge. Komentari korisnika na *online* platformama predstavljaju bogat izvor podataka o njihovim stavovima i mišljenjima, što pruža mogućnost za dublju analizu i razumevanje ponašanja potrošača.

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

Problem koji ovaj rad adresira je predikcija ocena (rejtinga) komentara korisnika na osnovu sentimenta izraženog u tekstualnom sadržaju tih komentara. Specifično, cilj je razviti modele koji mogu automatski analizirati tekstualne komentare i dodeliti im ocenu u rasponu od 1 do 5. Ovaj problem je izazovan zbog prirode tekstualnih podataka, koji su često neuredni, neformalni i kontekstualno zavisni.

Koriste se različiti pristupi dubokog učenja i NLP-a, uključujući RNN, CNN, *Word2Vec*, *GloVe*, BERT i generativni model *llama-7b-hf*. Podaci su preuzeti sa *Kaggle* platforme i uključuju 23,486 komentara korisnika [1]. Pretprocesiranje podataka obuhvata čišćenje teksta i pripremu za treniranje modela. Skup podataka je podeljen na trening, validacioni i testni deo za evaluaciju performansi modela.

Evaluacija je izvršena korišćenjem F1 mere. Najbolji rezultati postignuti su upotrebom transformera i SVM klasifikatora, dok su generativni modeli pokazali potencijal za dodatna istraživanja. Ovaj rad doprinosi unapređenju metoda za automatsku analizu sentimenta u komentarima korisnika, što može biti korisno za *e-commerce* platforme i druge aplikacije koje se oslanjaju na korisničke povratne informacije.

U prvom poglavlju dat je uvod sa opisom problema. U drugom poglavlju opisana su prethodna rešenja i najrelevantniji radovi. Treće poglavlje opisuje metodologiju predikcije, četvro poglavlje eksperimente i njihove rezultate, a peto poglavlje daje zaključke rada.

2. PRETHODNA REŠENJA

Rad "Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network" [2], istražuje vezu između recenzija kupaca i preporuka proizvoda, fokusirajući se na e-trgovinu. Autori su koristili 22,621 javni komentar. Implementirana je bidirekciona rekurentna neuronska mreža sa LSTM, uz korišćenje *GloVe* reprezentacije za mapiranje reči u vektorski prostor. Rejting od 3 ili više tretira se kao pozitivna ocena, dok se rejting manji od 3 tretira kao negativna ocena. Postignuta je F1 mera od 0.93 za klasifikaciju sentimenta, ali je uočena pristrasnost zbog nebalansiranosti pozitivnih i negativnih komentara.

U radu "Sentiment Analysis using machine learning algorithms: online women clothing reviews" [3], autor koristi različite metode za klasifikaciju sentimenta.

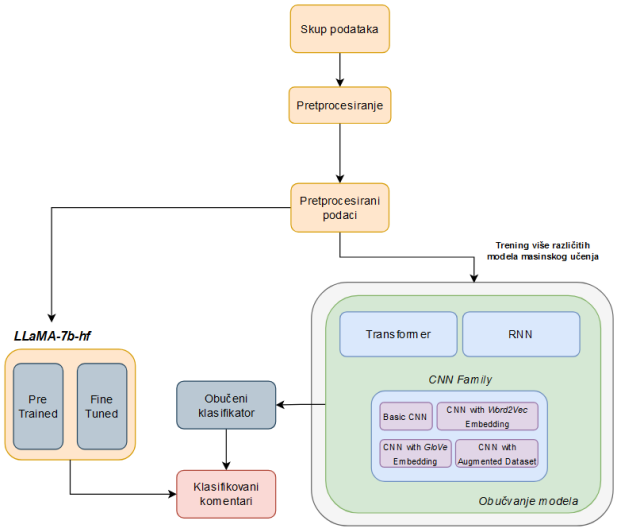
Korišćeni algoritmi uključuju *Support Vector Machine*, *Logistic Regression*, *Random Forest* i *Naive Bayes*. Modeli su evaluirani na osnovu tačnosti, preciznosti, odziva, F1 mere i AUC. *Naive Bayes* je postigao najveću preciznost od 93%.

Rad "Fine-grained Sentiment Classification using BERT" [4], predstavlja značajan doprinos u oblasti analize sentimenta, posebno kroz primenu *Bidirectional Encoder Representations from Transformers* (BERT) modela za precizno razlikovanje finih nijansi sentimenta. Eksperiment je izvršen nad *Stanford Sentiment Treebank fine-grained* (SST-5) skupom podataka, gde je BERT model postigao tačnost od 55,5%.

Rad "Self-Attention-Based BiLSTM Model for Short Text Fine-Grained Sentiment Classification" [5], uvodi model koji poboljšava analizu sentimenta kratkih tekstova fokusirajući se na nijanse i aspekte teksta. Model koristi BiLSTM i mehanizme samo-pažnje za bolje hvatanje konteksta i značaja reči. Evaluacija modela je izvršena nad tri skupa podataka sa *SemEval 2014 Task 4* takmičenja, gde je postignuta prosečna tačnost od 73% na prvom, 65% na drugom i 65% na trećem skupu podataka.

3. METODOLOGIJA

Metodologija korišćena u ovom radu za sentiment analizu komentara uključuje nekoliko ključnih koraka, koji su prikazani na slici 1.



Slika 1. Dijagram metodologije.

Prvo se primenjuje pretprocesiranje nad skupo podataka koji je preuzet sa sajta *Kaggle* i sadrži 23,486 redova u kojima su opisani komentari korisnika. Za potrebe projekta su korišćeni samo atributi 'Review Text' i 'Rating'. Ostala polja su izostavljena jer nisu relevantna za rešavanje problema kojim se ovaj rad bavi. Skup podataka je podeljen na trening, test i validacioni skup u razmeri 80/10/10, pri čemu trening skup sadrži 18,788 komentara, validacioni 2,349, a test skup 2,349 komentara. Podela je odrađena na osnovu atributa 'Rating', kako bi se očuvala ista distribucija ciljne labele u svakom od skupova podataka.

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

1. Uklanjanje praznih komentara koji ne sadrže korisne informacije
2. Izbacivanje duplikata kako bi se izbeglo višestruko učenje istih informacija
3. Zamena višestrukih razmaka (engl. *whitespace*) jednim razmakom za uniformnost teksta
4. Izbacivanje HTML tagova iz komentara koristeći *BeautifulSoup* [6] biblioteku
5. Zamena ili uklanjanje unicode karaktera koji mogu ometati procesiranje teksta
6. Uklanjanje nepoželjnih simbola koji mogu iskriviti analizu

Ove varijacije su implementirane kako bi se ispitalo koji pristup najbolje doprinosi poboljšanju performansi modela.

Pretprocesirani trening skup se koristi za obučavanje modela za klasifikaciju komentara. Pre nego što se tekstualni sadržaji mogu proslediti neuronskoj mreži za klasifikaciju, ključno je pretvoriti ih u vektore koji efikasno održavaju semantičku vrednost teksta. Dobijeni vektor teksta se zatim šalje na ulaz mreže. Implementacija konvolutivne mreže je odrađena upotrebom biblioteke *Keras* [7]. Tokenizacija teksta je realizovana korišćenjem *Keras*-ovog Tokenizera, dok su *Word2Vec* i *GloVe* korišćeni kao naprednije metode vektorizacije. Implementacija rekurentne mreže je odrađena upotrebom biblioteke *TensorFlow* [8]. Korišćena su dva bidirekcionalna LSTM sloja, svaki sa 16 neurona, i dodatni *Dropout* slojevi kako bi se sprečilo preterano prilagođavanje. Tokom treniranja, vrednost funkcije greške na validacionom setu je pažljivo praćena kako bi se osiguralo da model ne gubi sposobnost generalizacije na novim podacima. Da bi se ublažila prisutna nebalansiranost u *dataset*-u, primenjen je pristup augmentacije podataka korišćenjem tehnike parafraziranja koristeći PEGASUS model iz biblioteke *Transformers*. Korišćen je *nli-mpnet-base-v2* transformer model iz *sentence-transformers* biblioteke [9]. Pretrenirani transformer model korišćen je kao sloj za vektorizaciju teksta, dok je na izlaz ovog modela dodat klasifikator koji se trenira i daje konačnu predikciju rejtinga za dati komentar. Klasifikatori kao što su SVM, *K-Nearest Neighbors*, *Random Forest*, *Multi-Layer Perceptron*, *Logistic Regression* i *Linear Regression* su trenirani na trening skupu i evaluirani na validacionom skupu podataka. Za implementaciju generativnog pristupa korišćen je *llama-7b-hf* model [10]. Implementacija uključuje kreiranje konteksta zadatka za model, prikazivanje test primera i prikupljanje odgovora koje model generiše za konačnu klasifikaciju. Koriste se *AutoModelForCausalLM* i *SFTTrainer* klase za implementaciju *few-shot learning* metode, gde se modelu dostavljaju reprezentativni primeri za svaku klasu sa informacijama o klasama.

4. EKSPERIMENTI I REZULTATI

U ovom poglavlju istraženi su različiti pristupi sentiment analizi koristeći kombinaciju tradicionalnih i savremenih metoda obrade prirodnog jezika. Eksperimenti su

podeljeni u dve glavne grupe, poređenje klasičnih NLP pristupa i poređenje generativnog pristupa. Eksperimenti su obuhvatili evaluaciju efikasnosti modela na temelju detaljno pripremljenog skupa podataka, prilagođavanja arhitekture modela i njihovih hiperparametara, kao i primenu generativnih modela. Komparativna analiza performansi modela bazirana je na standardnim metrikama, kako bi se identifikovali najefikasniji pristupi za rešavanje problema sentiment analize.

Komentari o ženskoj odeći, koji su prošli kroz proces čišćenja i lematizacije koristeći biblioteku *Spacy* [11]. Zbog dominacije komentara sa ocenom 5, primenjena je augmentacija podataka za uravnoteženje skupa.

Eksperimenti u prvoj grupi su sprovedeni koristeći različite pristupe:

1. Konvolutivne neuronske mreže (CNN): Korišćenje osnovnog CNN modela, kao i varijanti sa *Word2Vec* i *GloVe* reprezentacijama reči
2. Rekurentne neuronske mreže (RNN): Upotreba bidirekcionalnih LSTM slojeva za obradu sekvencijalnih podataka
3. Transformer modeli: Implementacija BERT modela za finu analizu sentimenta

Tokom eksperimenata, optimizovani su hiperparametri i arhitekture modela korišćenjem grafičkih kartica dostupnih preko *Google Colab*-a. Proces treniranja modela uključivao je tehnike ranog zaustavljanja kako bi se sprečilo prekomerno učenje.

Eskperimenti iz druge grupe za generativni pristup **uključivali su testiranje llama-2-7b-chat-hf modela, u** varijantama sa i bez finog podešavanja. *Fine tuning* je omogućio prilagođavanje modela specifičnostima skupa podataka, dok je upotreba *few-shot learning*-a pomogla u klasifikaciji komentara.

Evaluacija performansi modela izvršena je korišćenjem F1 mere kao glavne metrike zbog nebalansiranosti skupa podataka. Najbolje rezultate postigao je transformer model u kombinaciji sa SVM klasifikatorom. Augmentacija podataka pokazala je minimalan uticaj na poboljšanje performansi.

Tabela 1. *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		0.6153
MLP *Classifier*	*n_estimators*=300 *max_depth*=10	0.6697
Logistic Regression CV	*alpha*=0.01, *early_stopping=True*	0.6644
Linear Regression	*multi_class='multinomial'*	0.6091

Tabela 2. *Rezultati nad skopom sa uklonjenim emotikonima*

Model	**F1 mera**
RNN	0.5298
CNN	0.5980
CNN + *GloVe pretrained embedding*	0.5979
CNN + *Augmented dataset*	0.5121
CNN + *Word2Vec embedding layer*	0.5592
Transformer + SVM	**0.6623**

Tabela 3. *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

Tabela 4. *Rezultati pretreniranog generativnog pristupa*

Klasa	**Preciznost**	**Odziv**	**F1 mera**	**Br. Primeraka**
Negative	0.23	0.99	0.37	300
Somewhat Negative	0,00	0,00	0,00	300
Neutral	0,19	0,02	0,03	300
Somewhat Positive	0,00	0,00	0,00	300
Positive	0,78	0,41	0,54	300
Tačnost			0,28	1500
Makro prosek	0,24	0,28	0,19	1500

Tabela 5. *Rezultati fine tuned verzije generativnog pristupa*

Klasa	**Preciznost**	**Odziv**	**F1 mera**	**Br. Primeraka**
Negative	0.26	0.95	0.40	300
Somewhat Negative	0.00	0.00	0.00	300
Neutral	0.43	0.08	0.13	300
Somewhat Positive	0.00	0.00	0.00	300
Positive	0.73	0.82	0.77	300
Tačnost			0.37	1500
Makro prosek	0.28	0.37	0.26	1500

Rezultati eksperimenata pokazali su da konvolutivne i transformer mreže pružaju bolje performanse u odnosu na rekurentne mreže. Generativni pristupi pokazali su potencijal, ali su njihovi rezultati bili varijabilni i zavisili su od specifičnih postavki modela i broja epoha treniranja. Klasifikacija komentara sa srednjim ocenama predstavljala je izazov zbog njihove inherentne neodređenosti, što ukazuje na potrebu za daljim istraživanjima i unapređenjima modela.

Ovi rezultati potvrđuju da savremeni modeli dubokog učenja, naročito transformeri, mogu značajno unaprediti tačnost i pouzdanost sentiment analize u *e-commerce* recenzijama. Dalje istraživanje i optimizacija ovih modela mogli bi dodatno poboljšati rezultate, posebno u scenarijima sa nebalansiranim skupovima podataka.

5. ZAKLJUČAK

Klasifikacija korisničkih komentara prema sentimentu predstavlja složen izazov koji može znatno poboljšati interpretaciju i interakciju s korisnicima kada se adekvatno adresira. U ovom istraživanju analizirani su različiti metodološki pristupi za sentiment analizu komentara prikupljenih s različitih platformi za prodaju ženske odeće, preuzetih s veb stranice *Kaggle*. Podaci su očišćeni, obrađeni i kategorizovani u pet grupa za klasifikaciju u više klasa.

Skup podataka je podeljen na obučavajući, validacioni i testni segment. Zbog nebalansiranosti skupa podataka, korišćene su makro usrednjene metrike preciznosti, osetljivosti i F1 mere za ocenu modela.

U eksperimentima su testirani različiti pristupi:

1. Tradicionalni NLP metodi: Tekst je vektorizovan pomoću *Word2Vec*, *GloVe*, LSTM, CNN i Transformer tehnika
2. Augmentacija podataka: Korišćena za rešavanje problema nebalansiranosti skupa podataka
3. Generativni modeli: Klasifikacija pomoću *llama-7b-hf* modela, u pretreniranoj i *fine tuned* varijanti

Najbolje rezultate postigao je Transformer model sa F1 merom od 0.66, dok su konvolutivne mreže pokazale nešto slabije rezultate. Generativni model, iako nije pravljen za klasifikaciju, pokazao je sposobnost za ovu namenu. Postoji pretpostavka da generativni modeli mogu biti efikasni za klasifikaciju u kontekstu manjih skupova podataka zahvaljujući svom opštem znanju.

Ograničenja u resursima sprečila su dalja istraživanja i korišćenje modela sa većim brojem parametara. Iako su generativni modeli napredovali, klasični NLP pristupi su se pokazali efikasnima kada postoji dovoljna količina podataka za obuku.

Dalji razvoj rešenja bi se fokusirao na prikupljanje šireg seta podataka, posebno komentara s nižim ocenama, kako bi se adresirala nebalansiranost *dataset*-a i omogućila bolja generalizacija modela. Ovo bi uključivalo primenu LSTM mreža i CNN-a sa *Word2Vec* i *GloVe* vektorima, kao i usavršavanje BERT modela za detaljniju analizu konteksta i sentimenta. Razvoj raznolikijih *prompt*-ova za generativne modele i testiranje modela sa više parametara mogli bi dodatno poboljšati efikasnost.

Ovaj pristup bi omogućio modelima bolje razumevanje i interpretaciju suptilnih ili kompleksnih varijacija sentimenta, naročito u nepolarizovanim komentarima. Kroz kolaboraciju sa industrijskim partnerima i akademskim institucijama, moguće je dobiti pristup naprednijim tehnološkim resursima i znanju, što bi dodatno doprinelo preciznosti i efikasnosti modela. *classification*

6. LITERATURA

[1] https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews (pristupljeno u martu 2022.)

[2] Abien Fred M. Agarap, „Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network“, 2018.

[3] Shuangyin Xie, „Sentiment Analysis using machine learning algorithms: online women clothin reviews“, 2019.

[4] Manish Munikar, Sushil Shakya, Aakash Shrestha, “Fine-grained Sentiment Classification using BERT”, 2019.

[5] Jun Xie, Bo Chen, Xinglong Gu, Fengmei Liang, Xinying Xu, “Self-Attention-Based BiLSTM Model for Short Text Fine-Grained Sentiment Classification”, 2018.

[6] https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (pristupljeno u martu 2022.)

[7] https://keras.io/ (pristupljeno u martu 2022.)

[8] https://www.tensorflow.org/ (pristupljeno u martu 2022.)

[9] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu, “MPNet: Masked and Permuted Pre-training for Language Understanding”, 2020.

[10] https://huggingface.co/meta-llama/Llama-2-7b-hf (pristupljeno u februaru 2024.)

[11] https://spacy.io/ (pristupljeno u martu 2022.)

Kratka biografija:

Cvijetin Mlađenović rođen je u Šapcu 1998. god. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva – Inteligentni sistemi odbranio je 2024.god.
kontakt: mladjenovic.cvijetin@gmail.com