Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 4.89
DOI: https://doi.org/10.24867/30BE01Jovic

KOMPARATIVNA ANALIZA TEKSTUALNIH MODELA BERT, BART I XLNET ZA PREDIKCIJU POPULARNOSTI YOUTUBE SNIMKA

COMPARATIV ANALYSIS OF TEXTUAL MODELS BERT, BART AND XLNET FOR PREDICTING THE POPULARITY OF YOUTUBE VIDEOS

Nela Jović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – Elektrotehnika i računarstvo

Kratak sadržaj – *Ovaj rad se bavi komparativnom analizom tri tekstualna modela BERT, BART i XLNet. Podaci za treniranje i validaciju su preuzeti sa sajta kaggle.com. Postoje dva skupa podataka. Prvi inicijalni i drugi proširen, koji je nastao zbog pokušaja poboljšanja recall metrike. Nakon preuzimanja urađeno je pretprocesiranje odnosno tokenizacija i stemovanje. Skup je podeljen u razmeri 80:20 za treniranje i validaciju. Isprobane su različite vrednosti za learning rate i batch size, a najbolji rezultat je dobijen korišćenjem BERT large modela kada je learning rate 1e-5, a batch size 8. Druga dva modela dala su nešto lošiji rezultat od BERT-a. Svi eksperimenti i rezultati biće predstavljeni u nastavku.*

Ključne reči: *BERT, BART, XLNet, Fine-tuning, Youtube*

Abstract – *This paper talks about a comparative analysis of three text models: BERT, BART, and XLNet. The data for training and validation were taken from the website kaggle.com. There are two datasets: the initial one and an expanded one, created in an attempt to improve the recall metric. After downloading, preprocessing was done, including tokenization and stemming. The dataset was split in a ratio of 80:20 for training and validation. Different values for learning rate and batch size were tested, and the best result was achieved using the BERT large model with a learning rate of 1e-5 and a batch size of 8. The other two models produced somewhat worse results than BERT. All experiments and results will be presented below.*

Keywords: : *BERT, BART, XLNet, Fine-tuning, Youtube*

1. UVOD

Svedoci smo koliko se savremenih internet zanimanja razvilo u poslednjih par godina. Sve više ljudi u svoju biografiju dodaje i posao influensera kao izvor dodatne zarade ili čak primarno zanimanje. Platforma koju mnogi koriste kada žele da se bave ovim poslom je upravo youtube. Na njoj je moguće objavljivanje snimaka različitog vremenskog trajanja i sadržaja. Moderni influenseri su u trci za što većom popularnošću kako bi sa njom i njihova zarada porasla.

Analizirajući tekstualne podatke o svakom snimku, modeli će pokušati da dođu do zaključka šta to jedan snimak čini

popularnim na ovoj platformi. Koliko se na osnovu naslova snimka, naziva autora, tagova i deskripcije može pretpostaviti da li će se snimak dopasti ljudima koji ovu platformu koriste da ispune svakodnevno vreme.

Problem koji se rešava je multiklasna klasifikacija. Za ovaj problem iskorišćena je fine-tuning tehnika nad već postojećim jezičkim modelima BERT, BART i XLNet. Ova tri modela koriste različite tehnike za rad sa tekstom i iz tog razloga će se porediti koja od postojećih tehnika najbolje rešava ovaj problem.

U svrhu ovog istraživanja trenirani su i validirani i veliki i mali modeli. Korišćena su dva skupa podataka tako što je svaki podeljen u razmeri 80:20 za trening i validaciju.

Najbolje se pokazao BERT large jezički model sa čak 69% tačnosti. Rezultate koje su XLNet i BART dali nisu mnogo lošiji od BERT-a, ali je kod njih bilo više naznaka da počinju da overfituju nego kod BERT-a.

Prvo poglavlje daje neki opšti uvid u rad, koji je problem i kako će biti rešen. Drugo poglavlje predstavlja uvid u radove koji se bave sličnim problemom i koriste slične tehnologije za rad. Potom, u trećem se vidi koja je to korišćena metodologija, kakvi su podaci, šta podrazumeva pretprocesiranje. Četvrto poglavlje je uvid u eksperimente i rezultate sa kratkom diskusijom i na kraju zaključkom koji je proizašao iz ovog rada.

2. PRETHODNA REŠENJA

U radu [1] korišćena je tehnika fine-tuning XLNet modela za detekciju lažnih vesti na društvenim mrežama. Skup podataka LIAR [2] sadrži oko 12800 vesti. Skup je podeljen na trening (10269), validacioni (1284) i test skup (1283). Od ovog skupa napravljena su dva klasifikaciona problema. Prvi je višeklasni gde postoji šest različitih kategorija, a to su istinite, neistinite, skoro istinite, polu istinite, uglavnom istinite i potpuno neistinite. Drugi problem je binarne prirode, gde postoje samo dve klase, istinite i lažne vesti. U ovom skupu podataka nalaze se podaci: ko je autor, koja je tema, država, posao, zabava, kakav je kredibilitet i koliko je istinita. Modelu je dodat izlazni sloj kako bi izračunao verovatnoću za svaku kategoriju. Najveća vrednost neke kategorije predstavlja konačan rezultat klasifikacije. Model se trenirao pet epoha, batch size je bio veličine 32 i korišćen je AdamW optimizator sa 1e-5 za learning rate vrednost. XLNet je

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

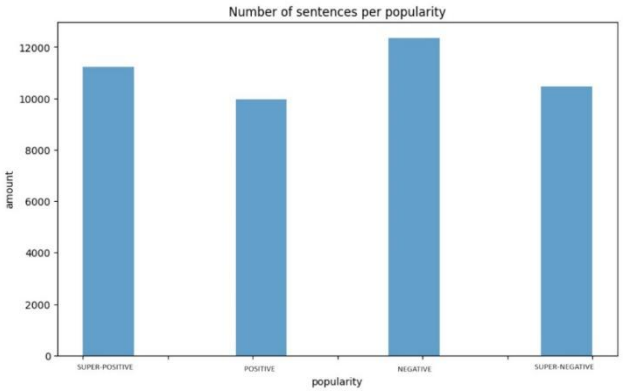
dao tačnost od 71%, BERT je za isti problem i podatke dao 62% za binarnu klasifikaciju, dok za multiklasnu klasifikaciju XLNet daje tačnost od 44%.

U radu [3] studija uvodi model za automatsku procenu povratnih informacija koristeći BERT jezički model. Model je treniran na skupu podataka od 10000 povratnih informacija koje su označene kao dobre ili loše. Dalje usavršavanje modela rađeno je na skupu podataka OULA. Model je dostigao tačnost od 93,4% na validacionom skupu, što ukazuje na njegovu sposobnost da proceni kvalitet povratnih informacija. Što se tiče pretprocesiranja rađeno je čišćenje, tokenizacija, feature extracting odnosno transformacija teksta u sekvencu embedinga koji sadrže semantičko značenje. BERT je poređen sa *Support Vector Machine (SVM)* algoritmom, *Random Forest, K-Nearest Neighbors, Logistic Regression*, Naivnim bajesom, stablom odlučivanja, Konvolutivinim neuronskim mrežama. BERT je dao najbolje rezultate po svim korišćenim metrikama (*F1, Recall, Precision*). Takođe, upoređen je i sa sebi sličnim modelima *RoBERT-a, PARsBERT-a* i *SemBERT-a* gde je dao najbolje rezultate.

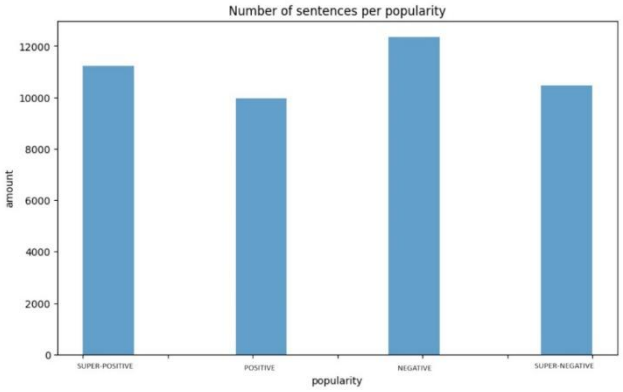
U radu [4] poredila su se tri modela : BERT, GPT i BART nad dva skupa podataka. Prvi je formiran od objava na twitter društvenoj mreži, a drugi od komentara na facebook-u. Prvo su testirani modeli bez treniranja, a zatim i posle treninga nad delom podataka. Prvi zadatak je bio da na osnovu twitter objave model zaključi kome je objava namenjena, Trampu ili Klintonovoj i da li je za, protiv ili neutralan. BART je pokazao najlošiji rezultat i najmanji F1 score dok je GPT dao najbolji rezultat. Slične rezultate dobili su i nakon treniranja. Čak je primećeno da je posle treninga narušena generalizacija koju BART ima.

3. FORMIRANJE SKUPA PODATAKA

Podaci su preuzeti sa sajta kaggle.com. Početni skup imao je 245987 podataka i ažurira se na dnevnom nivou. Kada su svi duplikati izbačeni ostalo je 44009 podataka sa jedinstvenim id-ijem. To predstavlja manji skup podataka nad kojim su se modeli trenirali i validirali. Veći skup nastao je proširivanjem prethodno pomenutog u cilju poboljšanja *recall* metrike. Veći skup podataka ima 67249 podatka i takođe su svi prikupljeni sa sajta *kaggle.com*. Kolone koje su korišćene su naziv snimka, naziv kanala, deskripcija i tagovi kao ulazni parametar, a labela koju je trebalo prediktovati je formirana na osnovu lajkova i dislajkova. Procenat broja dislajkova u odnosu na lajkove definisala je popularnost snimka. Naime, ako je procenat do 2% to je *super positive* klasa odnosno jako popularni videi. Zatim, od 2 do 5% snimci su *positive* odnosno popularni, od 5 do 10% *negative* odnosno nepopularni i preko 10% *super negative* odnosno veoma nepopularni. Na slikama 1. i 2. vidi se broj video snimaka po svakoj klasi. Naime, veći skup nastao je proširivanjem samo onim podacima koji čine *positive* i *negative* klasu jer je za njih uočen niži *recall* u odnosu na *super positive* i *super negative* klasu.



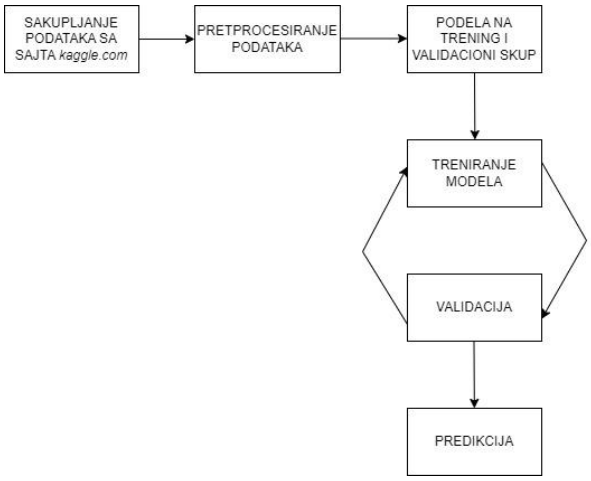
Slika 1. Broj video snimaka po klasama – mali skup



Slika 2. Broj video snimaka po klasama – veći skup

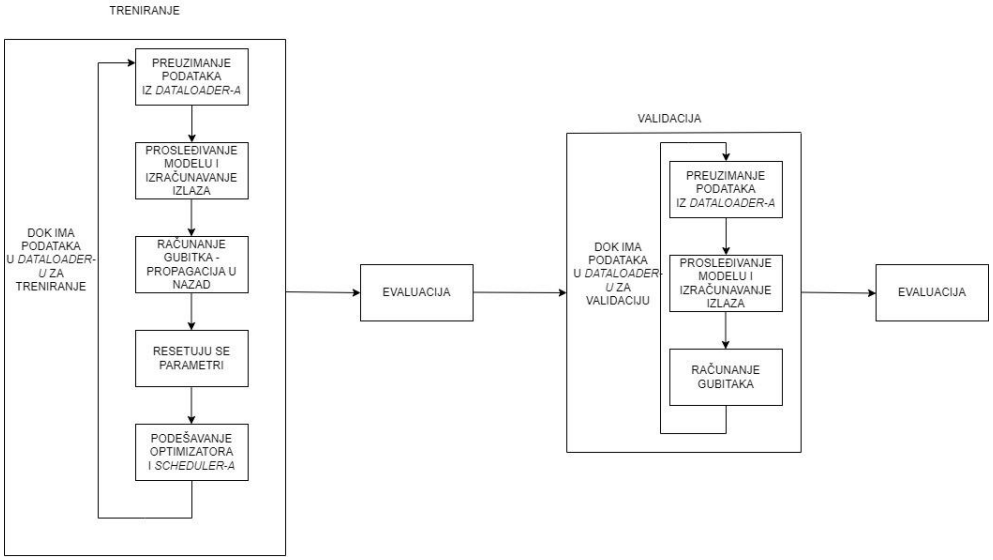
4. METODOLOGIJA

Dijagram na slici 3. prikazuje metodologiju rada.



Slika 3. Dijagram metodologije

Nakon dobavljanja podataka radi se pretprocesiranje koje uključuje tokenizaciju i stemovanje. Tako dobijeni skup podataka deli se u odnosu 80:20 na trening i validacioni skup. Definiše se *AdamW* optimizator kao i *scheduler*. Vrednosti *learning rate* i *batch size* su menjane tokom eksperimenata. Inicijalno su bile definisane četiri epohe u kojoj su se obavljali i trening i validacija. Na kraju svake epohe ispisivan je izveštaj metrika *recall, F1, precision, accuracy* i *loss* i za trening i za validaciju. Na taj način kontolisano je koliko model napreduje u učenju kao i *overfitting*. Na slici 4. prikazano je kako izgleda jedna epoha.



Slika 4. Tok jedne epohe

5. EKSPERIMENTI I REZULTATI

Eksperimenti su grupisani po modelima koji su korišćeni. Za svaki model obavljena su četiri eksperimenta, gde svaki predstavlja jedan od modela (*large* ili *base*) i jedan od skupa podataka (veći ili manji).

Naime, najbolji se pokazo BERT *large* model sa manjim skupom podataka, ali slične performanse daje i BERT *base* za veći skup podataka. *Loss* funkcija tokom epoha opada. U tabeli 1. prikazani su rezultati po klasama u četvrtoj epohi za vrednosti parametara learning rate = 1e-5 i batch size = 8 kada se koristi BERT large model i manji skup podataka.

Tabela 1. Metrike po klasama BERT large model

TRENING			
	precision	recall	F1
super positive	0.65	0.66	0.66
positive	0.58	0.51	0.54
negative	0.64	0.67	0.66
super negative	0.78	0.79	0.78
Macro avg	0.66	0.66	0.66
VALIDACIJA			
super positive	0.64	0.75	0.69
positive	0.61	0.54	0.57
negative	0.70	0.66	0.68
super negative	0.80	0.80	0.80
Macro avg	0.69	0.69	0.69

Tačnost koju je taj model postigao je 0.69 što je ujedno najbolja tačnost celokupnog eksperimenta.

Nakon BERT-a ista tehnika sa istim podacima pokušana je i sa BART modelom. Naime ni *large* ni *base* nisu uspeli da prevaziđu BERT-ove rezultate. Veći skup podataka je pomogao da se rezultati poboljšaju, posebno kod large modela gde BART dostiže tačnost od 0.67. U tabeli 2. prikazani su rezultati BART *large* modela kada je korišćen veliki skup podataka.

Tabela 2. Metrike po klasama BART large model

TRENING			
	precision	recall	F1
super positive	0.65	0.81	0.72
positive	0.78	0.72	0.75
negative	0.75	0.65	0.70
super negative	0.77	0.85	0.81
Macro avg	0.74	0.76	0.74
VALIDACIJA			
super positive	0.55	0.73	0.63
positive	0.73	0.69	0.71
negative	0.69	0.57	0.62
super negative	0.70	0.75	0.73
Macro avg	0.67	0.69	0.67

Treća grupa eksperimenata uključuje *large* i *base* XLNet modele kao i iste skupove podataka koji su se koristili i za prethodne eksperimente. U tabeli 3. predstavljeni su najbolje ostvareni rezultati sa XLNet modelom, odnosno rezultati nakon četvrte epohe treniranja XLNet large modela sa velikim skupom podataka. Tačnost koja je postignuta iznosi 0.67.

Tabela 3. Metrike po klasama XLNet large modeli

TRENING			
	precision	recall	F1
super positive	0.68	0.82	0.75
positive	0.77	0.74	0.75
negative	0.77	0.68	0.72
super negative	0.80	0.87	0.83
Macro avg	0.76	0.78	0.76
VALIDACIJA			
super positive	0.57	0.70	0.63
positive	0.73	0.68	0.70
negative	0.68	0.59	0.63
super negative	0.69	0.77	0.73
Macro avg	0.67	0.69	0.67

U tabeli 4. nalaze se tačnosti svakog eksperimenta za svaki skup podataka. Na osnovu nje možemo uporediti koji model se kako pokazao.

Tabela 4. Tačnost po modelima za skupove podataka

MODEL	TAČNOST ZA MANJI SKUP	TAČNOST ZA VEĆI SKUP
BERT base	0.59	0.67
BERT large	**0.69**	0.66
BART base	0.26	0.22
BART large	0.24	0.21
XLNet base	0.26	0.29
XLNet large	0.26	0.19

Iako se iz tabele iznad vidi da najbolju tačnost daje BERT large za manji skup, potrebno je deset sati treniranja i L4 GPU, dok za manji base model i veći skup podataka potrebno je duplo manje vreme i T4 GPU, a tačnost nije mnogo manja. XLNet i BART dali su malo lošiji rezultat. BART ne zahteva dugo treniranje, odnsno najmanje vremena je bilo potrebno upravo za njegovo obučavanje. Supotrno se ispostavilo za XLNet model čije treniranje je trajalo i preko šesnaest sati.

Analizom podataka za koje BERT nije dao dobar rezultat uočeno je da se među dosta primera nalaze naslovi koji u sebi sadrže link do nekog web sajta ili drugog youtube videa.

6. ZAKLJUČAK

U ovom radu predstavljena je komparativna analiza velikih jezičkih modela na istom problemu sa istim skupovima podataka. Upoređeni su modeli koji imaju različitu arhitekturu na istom klasifikacionom problemu. Motivacija za rad proizilazi iz sve veće popularnosti influenserskog zanimanja i pitanja da li će u budućnosti postojati jedan virtuelni asistent kao savetnik kako napraviti popularan youtube snimak.

Svaki eksperiment je tekao slično. Prvo su učitani i filtrirani podaci, tako da duplikati budu izbačeni. Potom su pretprocesirani podaci što uključuje tokenizaciju i stemovanje. Skup podataka podeljen je na trening i validacioni i inicijalizovane su četiri epohe.

Na osnovu rezultata uočava se da je BERT jedini model koji je uspeo da savlada ovaj zadatak u poređenju sa ostala dva modela. Tačnost u najboljem slučaju je 0.69, dok druga dva modela jako malo zaostaju za njim.

Kada sumiramo sve eksperimente vidimo da BERT base model nad većim skupom podataka daje tačnost 0.67 dok BERT large sa manjim skupom podataka daje 0.69. Treba uzeti u obzir da za base model treba duplo manje vremena i dovoljan je manji GPU, pa je zbog toga možda on optimalnije rešenje.

7. LITERATURA

[1] Kumar, Ashok; Trueman, Tina Esther; Cambria, Erik. Fake news detection using XLNet fine-tuning model. In: *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*. IEEE, 2021. p. 1-4.

[2] W. Y. Wang, „Liar, liar, pants on fire“ : A new benchmark dataset for fake news detection, *in Proceedings of 55th Annual Meeting of Association for Computational Linguistics* 2017, pp 2931-2937

[3] Spristav, Gaurav; Kant, Shri; Spristava, Durgesh. Design of an AI-Driven Feedback and Decision Analysis in Online Learning with Google BERT. *International Journal of Intelligent Systems and Applications in Engineering*, 2024, 12.10s: 629–643-629–643

[4] Chae, Youngjin; Davidson, Thomas. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*, 2023.

Kratka biografija:

Nela Jović rođena je u Zvorniku 2000. god. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva – Energetska elektronika i električne mašine odbranila je 2024.god.
kontakt: nelanekijovic@gmail.com