Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/22BE18Milunovic

KATEGORIZACIJA I ANALIZA SENTIMENTA DOKUMENATA UPOTREBOM VIŠEJEZIČNOG TRANSFORMER MODELA SA VIŠE IZLAZA

DOCUMENT CATEGORIZATION AND SENTIMENT ANALYSIS USING A MULTILINGUAL TRANSFORMER MODEL WITH MULTIPLE OUTPUTS

Dušan Milunović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – PRIMENJENE RAČUNARSKE NAUKE I INFORMATIKA

Kratak sadržaj – *U radu je predstavljen model za kategorizaciju i analizu sentimenta teksta na sto jezika upotrebom transformer neuronskih mreža. Takođe je pokazan način za optimizaciju takvog modela postupkom destilacije modela***.**

Ključne reči: *Sistemi za analizu teksta, klasifikacija teksta, sentiment analiza teksta, višejezični transformer modeli, modeli sa više izlaza*

Abstract – *This paper presents a model for categorization and sentiment analysis of texts in one hundred languages using transformer neural networks. A way to optimize that model using model distillation is also described.*

Keywords: *Natural language processing, text classification, sentiment analysis, multilingual transformer models, multi-output models*

1. UVOD

Oblast obrade prirodnog jezika (*Natural language processing – NLP*) je podoblast lingvistike, računarske nauke i veštačke inteligencije koja se bavi analizom teksta kako bi se iz njega izvukli različiti zaključci. U ovom radu je opisan model mašinskog učenja koji iz teksta može da izvuče sentiment i odredi kategoriju teksta. Sentiment predstavlja „ton“ teksta, koji može biti negativan, neutralan ili pozitivan, a određen je brojem između -1 i 1. Skup kategorija teksta izvučen je iz *GCP* taksonomije [1]. Izabran je drugi nivo detaljnosti koji sadrži 216 kategorija.

Dodatni zadaci i ograničenja ovog rada su bili da model treba dobro da radi na više jezika i da jedan model određuje i sentiment i kategoriju teksta.

Kako bi se ispunio cilj rada upotrebljen je transformer model *XLM-Roberta* [2], koji je pretreniran na sto jezika na *CommonCrawl* [14] skupu podataka. Upotrebom funkcionalnih modela *tensorflow* [12] biblioteke, kreira se jedan model koji rešava dva problema, kategorizaciju i analizu sentimenta teksta. To je značajno jer je veličina transformer modela njihov problem.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio prof. dr Platon Sovilj.

Na ovaj način, „telo” transformer neuronske mreže se kombinuje sa dve „glave” (glava za analizu sentimenta teksta i glava za kategorizaciju) i praktično se količina potrebnih resursa za rešavanje ova dva problema prepolovljava.
Poslednji problem kojim se bavi ovaj rad je dalja optimizacija modela postupkom destilacije, nalik na postupak predstavljen u radu istraživača *huggingface*-a [3].

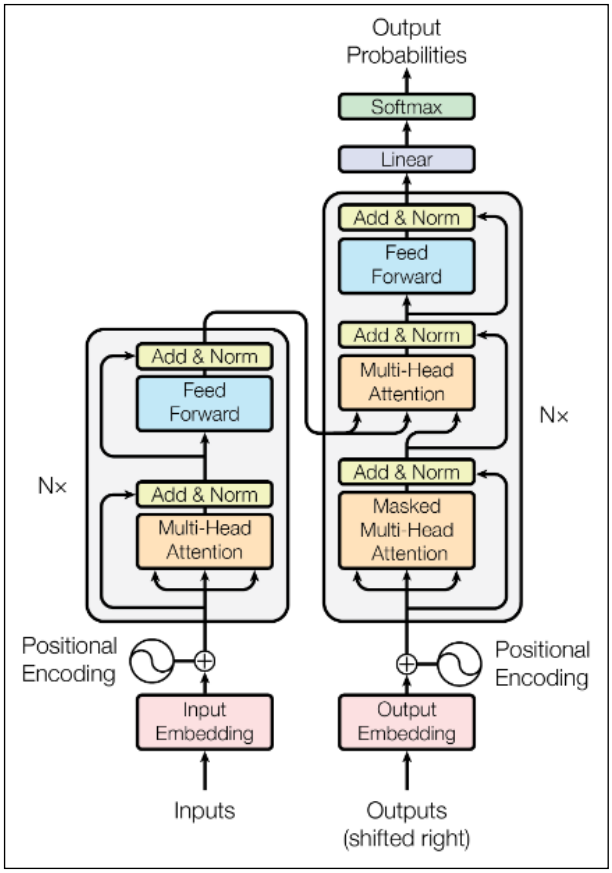
2. TEORIJSKE OSNOVE

2.1. Transformer neuronske mreže

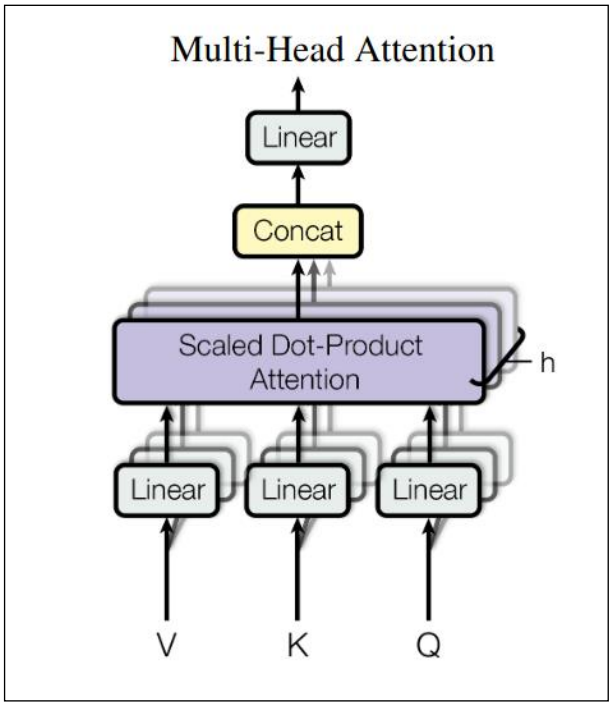
Transformer neuronske mreže [4] predstavljaju najuspešniju arhitekturu mreža za *NLP* u ovom momentu. Transformer neuronske mreže predstavljaju tip enkoder-dekoder mreža, čija je najčešća svrha da iz niza reči generiše drugi niz reči (npr. u slučaju prevođenja teksta). Za razliku od svojih prethodnika koji su koristili rekurentne neuronske mreže, transformeri izbacuju rekurentnu mrežu iz kombinacije i rade oslanjajući se samo na mehanizam pažnje. Rekurentne mreže imaju problem sa brzinom treniranja zbog svoje sekvencijalnosti, dok transformer mreže omogućavaju da se izračuna rezultat za čitav tekst jednim prolazom kroz mrežu. Ova činjenica omogućava brži trening transformer mreža i otvara vrata za treniranje nad većim skupovima podataka. Na slici 1 prikazana je arhitektura transformer neuronske mreže.

Sa leve strane slike nalazi se enkoder mreža, dok je sa desne strane dekoder. Za početak, vidi se da ulazni podaci i ciljani izlazni podaci prolaze kroz sloj za vektorizaciju (engl. *input/output embedding*). Taj sloj pretvara svaku reč u vektor fiksne dužine (standardne dužine su 128, 256, 512, 768 itd.). Višeglava pažnja (engl. *multi-head attention*) je glavna komponenta u ovoj mreži. Na slici 1. se vidi da se ona pojavljuje i u enkoder i u dekoder delu mreže.

Na slici 2. vidi se šematski prikaz višeglave pažnje. Ona se sastoji iz *h* ponovljenih komponenti, od kojih se svaka, u slučaju enkoder mreže, naziva samo-pažnja (engl. *self-attention*). Ulazi u mehanizam samo-pažnje su *V*, *K* i *Q*, koji redom označavaju vrednost (engl. *value*), ključ (engl. *key*) i upit (engl. *query*). Poenta mehanizma samo-pažnje je da odredi koji delovi ulaza se odnose jedni na druge. Dobar primer bi bio povezivanje zamenica sa imenicama koje ih predstavljaju, npr. u tekstu „Dečak je pomazio psa. To ga je učinilo srećnim”.



Slika 1. Arhitektura transformer mreže [4]



Slika 2. Višeglava pažnja [4]

Ne postoji jasna definicija na koga se reč „ga" odnosi. Mehanizam samo-pažnje mogao bi da pomogne tako što bi stvorio jaču vezu između reči „ga" i „dečak" ili „ga" i „psa"). Čitava poenta mehanizma samo-pažnje jeste da pametnije enkodira ulazne reči, uzimajući u obzir i ostale reči ulaza. Bitno je primetiti da se kod računanja pažnje taj račun može izvesti potpuno nezavisno za svaku reč iz ulaza. Ta osobina omogućava paralelizaciju treninga kod transformer modela, koja nije bila moguća kod rekurentnih mreža.

Na slici 1. vidi se da ulaz, osim što ulazi u sloj višeglave pažnje, takođe obilazi taj sloj takozvanom rezidualnom (engl. *residual*) vezom, nakon čega se sabira sa izlazom iz sloja višeglave pažnje i normalizuje. Normalizacija predstavlja način za regulisanje donjih i gornjih granica vrednosti u mreži i služi da se stabilizuje i ubrza trening. Nakon toga, rezultat ovih operacija prolazi kroz potpuno povezanu mrežu. Na isti način kao i ranije, ulaz u tu mrežu takođe ima rezidualnu vezu preko koje se sabira sa izlazom iz potpuno povezane mreže i opet se vrši normalizacija dobijenog zbira.

Nakon što se završi računanje vrednosti svih ulaznih reči u enkoderu, određuje se još jedan par vektora vrednosti V i ključeva K za svaku reč, koji služe kao ulaz u sloj višeglave pažnje u dekoder mreži. Da bi se dobili upiti Q za dekoder mrežu, koristi se mehanizam maskirane višeglave pažnje (engl. masked multi-head attention). Jedina razlika između obične i maskirane višeglave pažnje je što se u maskiranoj verziji, povezanost sa rečima koje u rečenici slede posle trenutno obrađivane reči postavlja na 0. Na taj način, sprečava se mogućnost da mreža posmatra reči koje još uvek nije generisala. Ovo je relevantno za vreme treninga transformer mreže, jer kasnije, tokom njene upotrebe, te reči zapravo neće biti unapred poznate. Izlaz iz maskirane višeglave pažnje služi kao vektor upita Q za sloj obične višeglave pažnje u dekoderu. Ostatak mreže radi na isti način kao i kod enkoder mreže. Nakon što poslednji u nizu dekodera izračuna svoj izlaz, taj izlaz prolazi kroz potpuno povezani sloj sa linearnom aktivacijom i na kraju kroz *softmax* sloj kako bi se odredilo koja će se reč generisati.

Poslednji koncept koji je uveden sa transformer mrežama je koncept pozicionog enkodiranja (engl. *positional encoding*) . U do sad objašnjenom delu transformer mreže ne postoji način da mreža uzme u obzir pozicije reči u rečenici. Poziciono enkodiranje rešava taj problem tako što dodaje (sabira) određen vektor vektoru svake ulazne reči. Vektori koji se dodaju zavise od pozicije reči u tekstu i sačinjavaju šablon koji mreža može da nauči. Intuicija iza ove tehnike je da se dodavanjem ovih vrednosti u vektore reči dodaje dovoljna razlika između njih da mreža može da nauči da ih razlikuje po poziciji u rečenici.

2.2. BERT i XLM-Roberta

BERT [5] (*Bidirectional Encoder Representations from Transformers*) predstavlja transformer neuronske mrežu, uz nekoliko velikih promena u odnosu na originalnu transformer mreža. Glavna je ta što je *BERT* samo enkoder deo transformer mreže. To znači da umesto dekodera, kao poslednji sloj *BERT*-a može da se izabere neuronska mreža za rešavanje proizvoljnog zadatka. Druga značajna razlika je što je *BERT* pretreniran model, koji služi za učenje prenošenjem znanja (engl. *tranfer learning*).

XLM-Roberta [2] je transformer model sličnih osobina kao *BERT*, ali sa ciljem da se koristi kao jedan model koji radi sa sto jezika. Stoga, on je pretreniran na sto jezika,

koristeći *CommonCrawl* program za prikupljanje 2.5 terabajta podataka. Još jedan od glavnih ciljeva ove mreže je postići mogućnost da se trenira za specifičan zadatak na jednom jeziku, a da se nakon toga koristi za isti zadatak na preostalim jezicima. Taj cilj se naziva međujezično razumevanje (engl. *cross-lingual understanding*) i *XLM-Roberta* trenutno predstavlja najbolje rešenje na tom polju.

2.3. Destilacija modela

Pojam destilacije modela je predstavljen u radu [6], gde je osnovna ideja bila da se znanje ansambla modela sakupi u jedan model, koji bi oponašao izlaze čitavog ansambla. Slična ideja je primenjena u polju *NLP* nad *BERT* modelom u radu [3]. Ovde, ideja je bila da se napravi model koji se sastoji od istih delova kao i *BERT*, ali sa manjim brojem slojeva. Rezultat rada je *DistilBERT* model, koji sadrži 40% parametara *BERT* modela, 60% je brži i postiže 97% performansi *BERT*-a na *GLUE Benchmark* [7] zadacima.

3. SKUP PODATAKA

Za analizu sentimenata iskorišćen je otvoren skup podataka sa projekta „*The Stanford Analysis Treebank*" [8]. On predstavlja skup ocena filmova sa označenim sentimentom.

S obzirom da otvoren skup podataka koji navodi kategoriju teksta nije pronađen, korišćena je tehnika pretraživanja interneta za tekstove koji odgovaraju kategorijama. Za početak, izabran je *GCP* (*Google Cloud Platform*) skup kategorija teksta. On sadrži tri nivoa kategorija po nivou opštosti. Za ovaj rad izabran je drugi nivo opštosti, koji sadrži 216 različitih kategorija. Zatim, izabrane su ključne reči za svaku od kategorija. Nakon toga, korišćenjem biblioteke *mediawiki* [9] pretražena je onlajn enciklopedija *Wikipedia* [10] za svaku ključnu reč i za svaku je pronađeno po 500 najrelevantnijih tekstova. S obzirom da je za slične ključne reči moguće dobiti iste linkove, izbačeni su duplikati. Koristeći istu biblioteku dobavljeni su sadržaji članaka sa *Wikipedia*-e. Time je dobijen skup podataka sa tekstovima na engleskom jeziku sa označenim kategorijama sačinjen od 224.879 primera. Ovaj način prikupljanja podataka je iskorišćen i opisan u radu [15].

4. IMPLEMENTACIJA REŠENJA

Za implementaciju *XLM-Roberta* neuronske mreže upotrebljena je kombinacija biblioteka *huggingface* [11] i *tensorflow* [12]. Tensorflow predstavlja radni okvir za kreiranje, treniranje i upotrebu neuronskih mreža u programskom jeziku *Python*. *Huggingface* je biblioteka specijalizovana za transformer neuronske mreže i sadrži veliki broj pretreniranih transformer mreža. *Huggingface* nudi implementacije modela u radnim okruženjima *tensorflow* i *pytorch* [13]. Za implementaciju *XLM-Roberta* modela kao osnova korišćena je klasa *TFXLMRobertaMode* biblioteke *huggingface*. Ona sadrži pretrenirano „telo" mreže i služi za dalje nadograđivanje za bilo koji *NLP* zadatak. Čitava mreža se može posmatrati kao jedan sloj koji služi za izvlačenje karakteristika teksta (engl. *feature extraction*).

Preostali deo mreže je deo za određivanje izlaza. Kako je potrebno odrediti i sentiment i kategoriju teksta, ovde su kreirane dve „glave" mreže. To je moguće implementirati funkcionalnim modelima *tensorflow*-a. Jedna od njih se bavi određivanjem sentimenta, a druga određivanjem kategorije teksta. Obe glave mreže su napravljene kao potpuno povezane mreže sa dodatkom sloja ispuštanja (engl. *dropout*).

Nakon treniranja *XLM-Roberta* modela, cilj je bio istrenirati manju mrežu koja će postizati slične rezultate koristeći destilaciju modela. Za studentsku mrežu izabran je model *DistilBERT-multilingual*. Razlog za izbor već postojeće mreže je što je treniranje potpuno novog studentskog modela postupak koji zahteva značajnu hardversku moć. Pre treniranja studentskog modela, napravljen je novi skup podataka, tako što je učiteljski model odredio distribucije verovatnoća za kategorije i sentiment za svaki tekst iz prvobitnog skupa podataka. Koristeći novi skup podataka i isti postupak treniranja kao kod učiteljskog modela, uspešno je istreniran studentski model. Rezultujući model sadrži oko 2.05 puta manje parametara od *XLM-Roberta* modela. Brzina novog modela je takođe oko 2 puta veća.

5. EVALUACIJA REŠENJA

Pri treningu modela primenjena je podela skupa podataka na trening, validacioni i testni skup podataka. Za trening podatke uzeto je 70% skupa podataka, za validacione 10%, a za testne 20% podataka. Rezultati u ovom odeljku se odnose na testni skup podataka, odnosno podatke sa kojima model nije radio ranije.

5.1. Evaluacija XLM-Roberta modela

Bitne metrike za treniranje *XLM-Roberta* modela pri treniranju se odvajaju na metrike vezane za kategorizaciju i metrike vezane za analizu sentimenta.

Za analizu sentimenta praćena je prosečna apsolutna greška (engl. *Mean absolute error*), odnosno apsolutna razlika između pravih i izračunatih sentimenata. Dobijena vrednost je 0.102.

Za klasifikaciju teksta korišćene metrike su mikro F-mera, tačnost i top 3 tačnost. Dobijena vrednost mikro F-mere je 0.596, tačnosti 60.1%, a top 3 tačnosti 78%.

U svrhu testiranja kako model radi na različitim jezicima, ručno je označeno 20 primera prikupljenih iz novinskih članaka na srpskom jeziku. Ti primeri su zatim prevedeni na engleski, nemački, španski i italijanski jezik upotrebom *Google Translate* alata. Cilj ovog dela evaluacije bio je dokazati da će model za sve, ili bar većinu jezika računati iste rezultate, bili oni većinom tačni ili većinom netačni. Za analizu sentimenta, vrednosti od -1 do 1 su podeljene u 5 kategorija (vrlo negativno, negativno, neutralno, pozitivno i vrlo pozitivno). Model je u 18 od 20 rezultata dao iste rezultate na svim jezicima. U jednom primeru je dao 4 od 5 istih rezultata, i u jednom 3 od 5 rezultata. Za kategorizaciju teksta posmatrano je da li model u svoja top 3 rezultata daje tačnu kategoriju (ukoliko daje, rezultat se smatra tačnim, a u suprotnom netačnim). Za 14 od 20 primera model je dao isti rezultat na svim jezicima. Za 5 primera je dao 4 od 5 istih rezultata, i u jednom 3 od 5 istih rezultata.

5.2. Evaluacija destilovanog modela

Za destilovani model, prvo su primenjene iste dve mere kao za *XLM-Roberta* model.

Prosečna apsolutna greška pri određivanju sentimenta sa destilovanim modelom je 0.135, što je za oko 0.033 lošije od originalnog modela. Mikro prosek F-mere *DistilBERT-multilingual*-a je 0.587, što je za oko 0.01 lošije od *XLM-Roberta*. Tačnost i top 3 tačnost destilovanog modela su 59.5% i 77.5%. Obe vrednosti su manje od vrednosti originalnog modela za otprilike 0.5%. Sve ove mere ukazuju na dobru uspešnost procesa destilacije.

Za testiranje modela na različitim jezicima, korišćeni su isti primeri kao sa originalnim modelom. Za analizu sentimenta, destilovani model u 15 od 20 primera daje isti rezultat na svim jezicima. U 2 primera daje 4 od 5 istih rezultata, a u preostala 3 primera daje 3 od 5 istih rezultata. Za kategorizaciju model u 8 od 20 primera daje isti rezultat za sve jezike. U 11 primera daje 4 od 5 istih rezultata, i u jednom primeru daje 3 od 5 istih rezultata. Ovo ukazuje da destilovani model nije u mogućnosti da potpuno replicira uspeh originalnog modela u radu sa različitim jezicima.

Poslednja mera koja se posmatrala kod destilovanog modela je koliko često on daje iste kategorije za tekst kao i originalni model. Izračunato je da u 73% slučajeva, oba modela daju potpuno isti rezultat. Ukoliko se posmatra da li se rezultat originalnog modela nalazi među najverovatnija tri rezultata destilovanog modela, dobija se da se u 91.5% slučajeva nalazi. Ove metrike pokazuju da destilovani model dobro oponaša originalni.

6. ZAKLJUČAK

U ovom radu predloženo je rešenje za problem kategorizacije i analizu sentimenta tekstova na sto različitih jezika upotrebom modernih transformer arhitektura.

Model je konstruisan uz oslonac na biblioteku *huggingface* kao osnovu za transformer modele i biblioteku *tensorflow* za prilagođene slojeve koji omogućuju da model ima dve izlazne vrednosti.

Najznačajniji doprinos ovog rada predstavljaju primenjene optimizacije nad transformer modelima. Prva je ta što se "telo" transformer mreže koristi za ekstrakciju karakteristika teksta, i njen izlaz se koristi kao ulaz za dve "glave" mreže koje određuju sentiment i kategoriju teksta.

Druga optimizacija jeste destilacija originalnog modela čija je osnova *XLM-Roberta* transformer model u manji model čija je osnova *DistilBERT-multilingual* model.

Obe optimizacije predstavljaju korak ka dovođenju transformer mreža u stanje u kom su dostupnije za upotrebu u realnom svetu, gde su performanse modela (brzina i zauzeće memorije) direktno vezane za cenu upotrebe i potrošnju električne energije modela.

7. LITERATURA

[1] *GCP* taksonomija https://cloud.google.com/natural-language/docs/categories (pristupljeno u septembru 2022.)

[2] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[3] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

[5] 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*.

[6] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[7] GLUE Benchmark https://gluebenchmark.com/

[8] The Stanford Analysis Treebank https://nlp.stanford.edu/sentiment/

[9] Mediawiki https://github.com/barrust/mediawiki

[10] Wikipedia https://www.wikipedia.org/

[11] Huggingface https://huggingface.co/

[12] Tensorflow https://www.tensorflow.org/

[13] Pytorch https://pytorch.org/

[14] CommonCrawl https://commoncrawl.org/

[15] Agarwal, A., Dahleh, M., Shah, D., Sleeper, D., Tsai, A., & Wong, M. (2019). Zorro: A Model Agnostic System to Price Consumer Data. *arXiv preprint arXiv:1906.02420*.

Kratka biografija:

Dušan Milunović rođen je u Novom Sadu 1997. god. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva – Primenjene računarske nauke i informatika odbranio je 2022.god.
kontakt:
dusanmilunovic17@gmail.com