



UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA
U NOVOM SADU




Marko Radović

**POREĐENJE MODELA ZA
IDENTIFIKACIJU SEMANTIČKE
SLIČNOSTI REČENICA**

DIPLOMSKI RAD
-Osnovne akademske studije-

Novi Sad, 2021.

	Univerzitet u Novom Sadu FAKULTET TEHNIČKIH NAUKA 21000 NOVI SAD, Trg Dositeja Obradovića 6	Datum:
	ZADATAK ZA IZRADU DIPLOMSKOG (BACHELOR) RADA	List:
		1/1

(Podatke unosi predmetni nastavnik - mentor)

Vrsta studija:	Osnovne akademske studije
Studijski program:	Softversko inženjerstvo i informacione tehnologije
Rukovodilac studijskog programa:	vanr. prof. dr Miroslav Zarić

Student:	Marko Radović	Broj indeksa:	SW 30/2015
Oblast:	Mašinsko učenje		
Mentor:	prof. dr Aleksandar Kovačević, vanredni profesor		

NA OSNOVU PODNETE PRIJAVE, PRILOŽENE DOKUMENTACIJE I ODREDBI STATUTA FAKULTETA IZDAJE SE ZADATAK ZA DIPLOMSKI RAD, SA SLEDEĆIM ELEMENTIMA:

- problem – tema rada;
- način rešavanja problema i način praktične provere rezultata rada, ako je takva provera neophodna;
- literatura

NASLOV DIPLOMSKOG (BACHELOR) RADA:

Poređenje modela za identifikaciju semantičke sličnosti rečenica

TEKST ZADATKA:

Zadatak ovog rada je evaluacija više pretreniranih *sentence embedding* modela u odnosu na više skupova podataka za semantičku sličnost i identifikaciju parafraze u cilju upoređivanja performansi modela u odnosu na skupove podataka.

Rukovodilac studijskog programa:	Mentor rada:

Primerak za: ☐ - Studenta; ☐ - Mentora

SADRŽAJ

1. UVOD	1
2. PRETHODNA REŠENJA	3
3. TEORIJSKI POJMOVI I DEFINICIJE	5
3.1. Veštačke neuronske mreže	5
3.2. Transformer neuronska mreža	7
3.2.1 <i>Input embedding</i> i <i>positional encoding</i>	8
3.2.2 <i>Self attention</i>	9
3.2.3 <i>Multi-head attention</i>	10
3.2.4 <i>Residual</i> veza i normalizacioni sloj	10
3.2.5 <i>Feed forward</i> sloj	10
3.3. <i>BERT (Bidirectional Encoder Representations from Transformers)</i>	11
4. SPECIFIKACIJA I IMPLEMENTACIJA REŠENJA	13
4.1. Korišćeni alati	13
4.2. Modeli neuronskih mreža	13
4.2.1 <i>Sentence transformers</i>	14
4.2.2 <i>Universal sentence encoder</i>	15
4.2.3 Kombinovani modeli neuronskih mreža	15
4.3. Skupovi podataka	16
4.3.1 Prva grupa skupova podataka	16
4.3.2 Druga grupa skupova podataka	17
5. METODOLOGIJA	23
6. REZULTATI I DISKUSIJA	25
6.1. Rezultati za prvu grupu skupova podataka	25
6.2. Rezultati i diskusija za drugu grupu skupova podataka	28
6.2.1 Analiza predikcija modela za <i>PARADE</i> skup podataka	31
7. ZAKLJUČAK	35
8. LITERATURA	37

1. UVOD

Identifikacija semantički sličnih tekstova (engl. *semantic textual similarity*, skraćeno *STS*) na nivou reči, rečenica, paragrafa i dokumenata ima bitnu ulogu u oblasti obrade prirodnog jezika (engl. *natural language processing*, skraćeno *NLP*). *STS* na nivou rečenica ima primenu u klasterovanju, semantičkoj pretrazi, detekciji duplikata, itd. U (Grootendorst, 2020) da bi se odredile teme u tekstu (engl. *topic modeling*) tekst se prevede u vektorsku reprezentaciju (engl. *embedding*), potom se smanji dimenzionalnost vektora, a nakon toga se primeni klasterovanje. Dobijeni klasteri predstavljaju grupe dokumenata koji su slični po temama koje se nalaze u njima. Problem pretrage po ključnim rečima je to što nije moguće izvući semantičko značenje upita. Primer semantičke pretrage je sistem u (Deshmukh i Sethi, 2020) gde autori koriste semantičko poklapanje upita sa traženim novinskim člankom. Dupliranje istih pitanja na forumima otežava pronalaženje odgovora, jer su odgovori na isto pitanje raspoređeni na više pitanja koja imaju isto značenje. Da bi sprečila gomilanje duplih pitanja na svom forumu, *Quora* (<https://www.quora.com/>) je organizovala takmičenje (<https://www.kaggle.com/c/quora-question-pairs>) gde je više timova pokušalo da reši ovaj problem.

Da bi se napravili uspešni *STS* sistemi, u većini slučajeva je neophodno posedovati veliku količinu anotiranih podataka, koji će biti upotrebljeni za treniranje nekog *STS* algoritma. Prikupljanje velike količine anotiranih podataka je novčano skup proces, vremenski zahtevan i neretko je potrebno angažovati domenske eksperte. U cilju ublažavanja ovog problema može se primeniti *transfer learning* (Raffel i drugi, 2019) gde je potrebno mnogo manje anotiranih podataka. U situaciji potpunog nedostatka anotiranih podataka, mogu se primeniti modeli koji su trenirani na nekom drugom tasku, a koji bi radili dobro na ciljnom tasku (*zero-shot learning*) ili modeli pretrenirani na velikoj količini teksta a koji su prilagođeni (engl. *fine-tuning*) nekom *STS* skupu podataka.

Cilj ovog rada je da se ispita da li u trenutnoj literaturi postoji model koji se može primeniti za *STS* i detekciju parafraze na nivou rečenica i koji radi zadovoljavajuće dobro za bilo koji domen i bilo koju vrstu teksta na engleskom jeziku, bez korišćenja anotiranih podataka. U tu svrhu modeli koji trenutno imaju najbolje performanse (engl. *state-of-the-art*, skraćeno *sota*) će biti evaluirani nad 15 različitih skupova

podataka kako bi se pokrio širok krug različitih vrsta tekstova dobavljenih sa društvenih mreža, iz akademske literature, web stranica, itd. a takođe i širok krug domena kao što su: medicina, računarske nauke, itd. Dobijeni rezultati će biti upoređeni sa modelima koji su trenirani nad ili prilagođeni svakom skupu podataka posebno, kako bi se odredilo koliko su modeli bez prilagođavanja daleko od performansi prilagođenih modela.

2. PRETHODNA REŠENJA

Za identifikaciju semantičke sličnosti na nivou rečenica, prva rešenja koriste sintaksnu sličnost upotrebom algoritma kao što je *Levenshtein Distance* (Levenshtein, 1966) ili leksičku sličnost kao u (Cordeiro i drugi, 2007). Ovi i drugi pristupi zasnovani na poklapanju *n-gram*-a, gramatičkoj strukturi rečenice, itd. imaju veliki nedostatak što ne rade dobro za rečenice koje su semantički slične, a koje sadrže različite reči. Da bi se nastavio napredak na ovom polju bilo je neophodno na neki način obučiti mašinu da razume značenje teksta.

Velika prekretnica u tom pravcu je ostvarena *Word2Vec* modelom (Mikolov i drugi, 2013), gde se svaka reč može predstaviti vektorom fiksne dužine, čije vrednosti “sadrže” semantičko značenje te reči. Postoje dve verzije ovog modela. Prva verzija (*CBOW*) je trenirana tako da na osnovu konteksta (nekoliko reči pre trenutne reči i isti broj reči nakon trenutne reči) predvidi trenutnu reč. Druga verzija (*Skip-gram*) je trenirana obrnuto, da na osnovu trenutne reči predvidi isti broj prethodnih i narednih reči. Da bi se ovaj model iskoristio za rečenice, neophodno je primeniti uprosečavanje, sabiranje ili neku drugu operaciju nad vektorima za svaku reč, kako bi se dobio vektor fiksne dužine koji će reprezentovati rečenicu. Problem sa *Word2Vec* modelom je taj što svaka reč ima uvek istu vektorsku reprezentaciju, a značenje reči može zavisi od konteksta u kojem se nalazi. Ovaj problem pokušavaju da reše modeli koji uzimaju u obzir kontekst u kojem se reč nalazi, kao na primer *ELMO* model (Peters i drugi, 2018). *ELMO* je *biLM* (*bidirectional language model*), treniran je da na osnovu prethodnih reči u rečenici predvidi narednu reč, a takođe je treniran u suprotnom smeru, da na osnovu narednih reči predvidi prethodnu reč. Ovaj način treniranja ne zahteva labelirane podatke, pa je pretreniran na velikoj količini teksta. Osnovu modela čine dve *LSTM* mreže (Hochreiter i Schmidhuber, 1997). Izlazi te dve mreže se množe sa odgovarajućim težinama a nakon toga se sabiraju, što čini konačan izlaz modela. Autori u (Ranasinghe i drugi, 2019) su pokazali kako upotrebom kontekstualizovanih modela i naprednijim algoritmima za formiranje rečenica od vektora reči, uspevaju da prevaziđu performanse *Word2Vec* modela. Posmatrajući *SICK* (Marelli i drugi, 2014) skup podataka i *pearson correlation coefficient* kao metriku, njihov model postiže performanse 0.753 dok *Word2Vec* model postiže 0.734.

Jedan od prvih uspešnih modela koji direktno enkoduju rečenicu je *Skip-Though* (Kiros i drugi, 2015). Model je nenadgledano obučavan tako da na osnovu ulazne rečenice rekonstruiše prethodnu i narednu rečenicu iz teksta. Može se upotrebiti ne samo za rečenice nego i za druge kraće tekstove kao što su paragrafi. Postiže 0.8655 *pearson correlation* na *SICK* skupu podataka. *InferSent* (Conneau i drugi, 2017) model je treniran na nadgledan način nad *Stanford Natural Language Inference (SNLI)* (Bowman i drugi, 2015) skupu podataka i postiže 0.885 *pearson correlation* na *SICK* skupu podataka, što je bolje od rezultata *Skip-Though* modela. Sledeći model koji se pokazao uspešniji od prethodnih za neke skupove podataka je *universal sentence encoder (USE)* (Cer i drugi, 2018). Ovaj model je prvo pretreniran nenadgledanim obučavanjem nad velikom količinom teksta sa interneta, a potom nadgledanim obučavanjem nad *SNLI* skupom podataka. U odnosu na *STS benchmark* (Cer i drugi, 2017) ovaj model postiže 0.766 *pearson correlation* dok *InferSent* postiže 0.758 (Cer i drugi, 2018). Poslednji pojedinačni model koji je značajno uticao na unapređenje performansi je *Sentence-BERT* (Reimers i Gurevych, 2019). Osnovu modela čini *BERT* (Devlin i drugi, 2018) model, a arhitektura modela je posebno pravljena kako bi se efikasno generisale kvalitetne vektorske reprezentacije rečenica. Postiže 0.792 *pearson correlation* za *STS benchmark* skup podataka podižući performanse u odnosu na *universal sentence encoder*. Dalja unapređenja ovog modela su moguća korišćenjem nekog drugog baznog modela osim *BERT*-a.

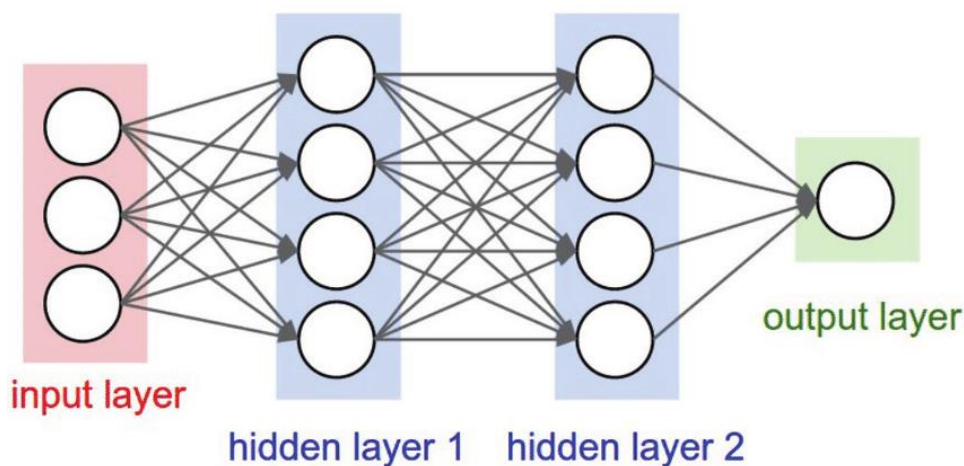
Kako bi se dalje podigle performanse, moguće je kombinovati više pojedinačnih modela. Autori u (Poerner i drugi, 2019) kombinuju *Sentence-BERT*, *universal sentence encoder* i *ParaNMT* (Wieting i Gimpel, 2017) modele i prevazilaze performanse svakog od ovih pojedinačnih modela, podižući najbolji rezultat za *STS benchmark* skup podataka na 0.839 *pearson correlation*.

3. TEORIJSKI POJMOVI I DEFINICIJE

Metod mašinskog učenja korišćen za konvertovanje rečenice prirodnog jezika u semantički razumljiv oblik za kompjuter u ovom radu je enkoder deo transformer neuronske mreže (Vaswani i drugi, 2017). U ovom poglavlju će prvo biti objašnjene klasične veštačke neuronske mreže (poglavlje 3.1) koje predstavljaju osnovu za razumevanje transformer neuronskih mreža predstavljenih u poglavlju 3.2 i u poglavlju 3.3 model koji je osnova za veći deo modela korišćenih u ovom radu.

3.1. Veštačke neuronske mreže

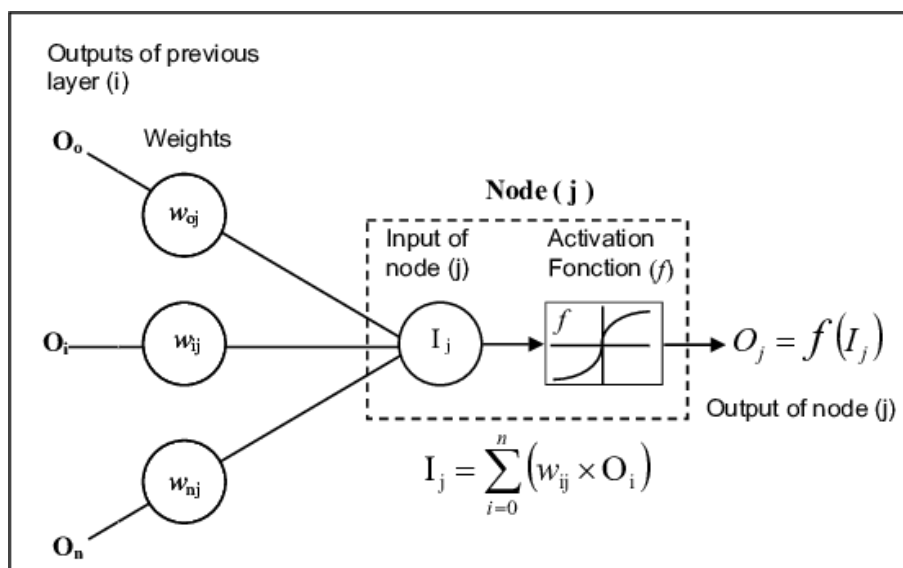
Veštačke neuronske mreže spadaju u algoritme nadgledanog obučavanja. Specifične su po tome što nije potrebno ručno odrediti obeležja za podatke (engl. *feature engineering*) već mreža to sama radi automatski. Cilj istreniranog modela je klasifikacija datog uzorka (ulaza), na osnovu nekih njegovih karakteristika, u jednu od predefinisanih klasa (izlaz).



Slika 1. Arhitektura neuronske mreže sa tri sloja (prvi sloj se ne broji)
(https://www.researchgate.net/figure/Topology-of-a-simple-neural-network_fig3_314224469)

Neuronska mreža je sačinjena od neurona i veza između njih. Neuroni su grupisani po slojevima. Prvi sloj, tj. ulazni, razlikuje se od ostalih slojeva po tome što mu se njegovi ulazi prosleđuju direktno iz samog uzorka. Što znači da broj neurona ulaznog sloja mora da odgovara dimenzijama ulaza. Izlaz poslednjeg sloja, tj. izlaznog sloja predstavlja rezultat algoritma. Broj neurona u izlaznom sloju se određuje u zavisnosti od rešavanog problema. Slojevi između ulaznog i izlaznog sloja se nazivaju skriveni slojevi (engl. *hidden layers*). Broj slojeva neuronske mreže se računa tako što se od ukupnog broja slojeva oduzme ulazni sloj. Svi neuroni iz jednog sloja su povezani sa svakim neuronom iz narednog sloja, tj. slojevi su međusobno potpuno povezani (engl. *fully connected layer*). Arhitektura opisanog modela je data na slici 1. Broj skrivenih slojeva u mreži i broj neurona u svakom sloju mreže predstavljaju osnovne hiperparametre modela.

Svaki čvor neuronske mreže predstavlja graf izračunavanja. Ako se izuzmu neuroni ulaznog sloja, ulazi neurona predstavljaju izlaze svih neurona iz prethodnog sloja. Kao što je prikazano na slici 2 svaki ulaz u neuron se množi sa odgovarajućom težinom. Težine se najčešće inicijalizuju sa nasumičnim malim brojevima a prilagođavaju se u procesu obučavanja. Nakon sabiranja svih proizvoda ulaza sa težinama rezultat se prosleđuje aktivacionoj funkciji. Aktivaciona funkcija uvodi nelinearnost i time omogućava rešavanje problema koji nisu linearno separabilni. Rezultat primene aktivacione funkcije predstavlja izlaz neurona.



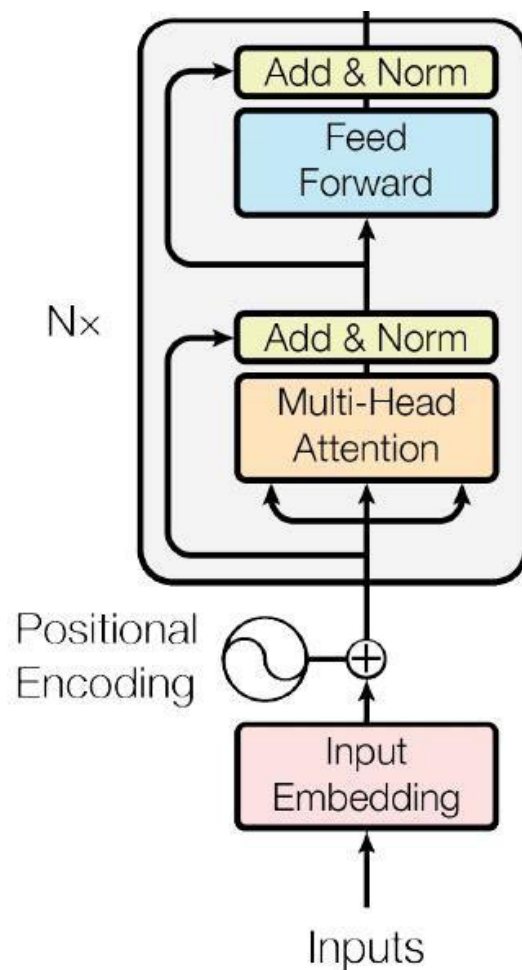
Slika 2. Veštački neuron

(https://www.researchgate.net/figure/The-basic-element-of-a-neural-network-node-computation_fig2_255629329)

Model neuronske mreže se trenira upotrebom *backpropagation* algoritma (Linnainmaa, 1970). Ukratko kada se jedan uzorak dovede na ulaz neuronske mreže, propagira kroz mrežu i izračuna izlaz (engl. *forward pass*) tada se izračuna greška algoritma, tj. razlika između stvarne vrednosti (engl. *ground truth label*) i prediktovane vrednosti. Potom se ta greška propagira kroz mrežu, od izlaznog sloja ka ulaznom sloju. Računanjem gradijenta dobija se koliko je svaki neuron uticao na grešku, pa se težine neurona ažuriraju u cilju smanjenja te greške.

3.2. Transformer neuronska mreža

Transformer neuronska mreža (Vaswani i drugi, 2017) nastaje kao rešenje na ograničenja do tada najzastupljenijih neuronskih mreža za rešavanje sekvencijalnih problema: rekurentnih mreža, *LSTM* (Hochreiter i Schmidhuber, 1997) i *GRU* (Chung i drugi, 2014). U radu je mreža trenirana na problemu prevoda sa jednog jezika na drugi, ali se može primeniti na mnoge druge probleme. U ovom radu se koriste modeli koji koriste samo enkoder deo originalnog transformer modela, iz tog razloga će biti izostavljeno objašnjenje dekodera dela. Arhitektura enkoder dela transformer mreže je prikazana na slici 3. U narednim poglavljima će biti detaljno objašnjeni delovi enkodera.



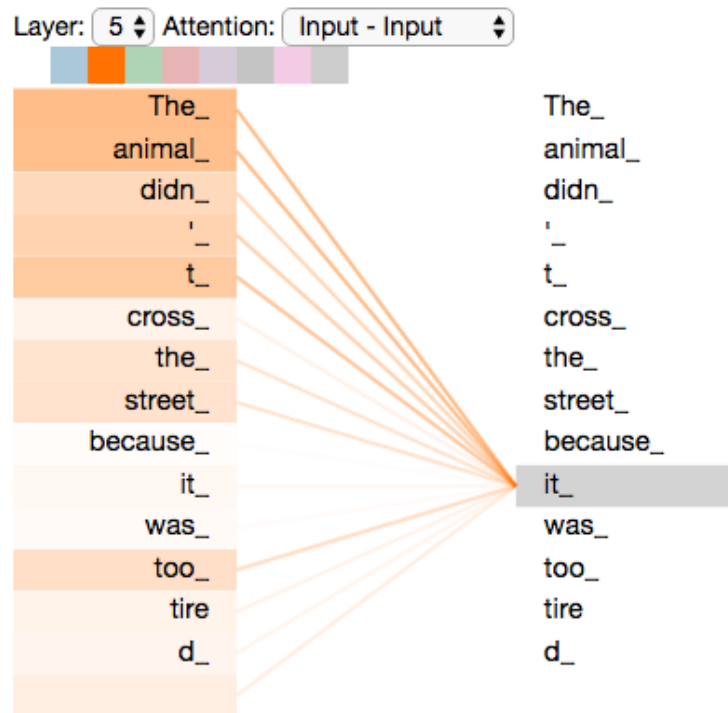
Slika 3. Enkoder deo transformer neuronske mreže
(<https://jalammar.github.io/illustrated-transformer/>)

3.2.1 *Input embedding i positional encoding*

Pre nego što se ulazni tekst prosledi u enkoder potrebno je konvertovati ga u vektorsku reprezentaciju. Tekst se najpre konvertuje u niz tokena, a potom tokeni u vektorske reprezentacije. Neki od često korišćenih algoritama za to su *BPE* (Sennrich i drugi, 2015), *WordPiece* (Wu i drugi, 2016), *SentencePiece* (Kudo i Richardson, 2018) itd. Da bi model uzeo u obzir poziciju reči u tekstu, na *input embedding* se dodaje *positional encoding*, vektorskim sabiranjem. U tu svrhu moguće je koristiti naučene *positional encoding* ili ih generisati upotrebom neke funkcije. Autori u radu koriste specijalnu formulu koja koristi sinusnu i kosinusnu funkciju u zavisnosti od pozicije tokena u tekstu.

3.2.2 Self attention

Posmatrajući sa višeg nivoa apstrakcije, ovaj princip omogućava modelu da za svaki token iz prosleđenog teksta identifikuje u kojoj meri su drugi tokeni iz istog tog teksta u vezi sa njim. Na taj način, model je u mogućnosti da kreira semantički bogatije vektorske reprezentacije za svaki token. Vizuelni prikaz na primeru jedne rečenice je dat na slici 4.



Slika 4. Vizuelizacija rezultata self attention mehanizma. Sa slike se vidi kako je model zaključio da se token 'it' odnosi na token 'animal'
(<https://jalammar.github.io/illustrated-transformer/>)

Za računanje *attention* skora, prvi korak je da se za svaki token izračunaju *query*, *key* i *value* matrice. One se računaju tako što se *input embedding*, na koji je dodat *positional encoding*, pomnoži sa tri matrice težina. Potom, ako se posmatra primer sa slike 4, za token 'it' se uzme *query* matrica, za token 'animal' *key* i *value* matrica, i ubaci se u formulu:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

gde su:

Q – *query* matrica,

K – *key* matrica,

V – *value* matrica,

d_k – dimenzija matrice Q .

Kada se izračuna attention skor za sve tokene u odnosu na token ‘*it*’, tada se svi skorovi saberu i dobija se konačni *self attention* skor za token ‘*it*’. Ovaj proces se primenjuje za svaki token posebno.

3.2.3 *Multi-head attention*

Multi-head attention predstavlja *self attention* ponovljen n puta za svaki token, gde je n broj glava. Svaka glava poseduje zasebne matrice težina za *query*, *key* i *value* matrice. Na taj način se za svaki token dobija n različitih reprezentacija koje se konkatenuiraju, pomnože sa dodatnom matricom težina koja se trenira i krajnji rezultat predstavlja *multi-head attention* skor. Prednost višestruke inicijalizacije težina je što će svaka krajnje istrenirana matrica uhvatiti različite zavisnosti između trenutno posmatranog tokena i ostalih tokena u prosleđenom tekstu.

3.2.4 *Residual veza i normalizacioni sloj*

Oko *multi-head attention* i *feed forward* sloja nalazi se *residual* veza. To znači da se tokom propagacije unapred, ulazi i izlazi sloja oko kojeg se nalazi residual veza sabiraju. Nakon *multi-head attention* i *feed forward sloja* nalazi se normalizacioni sloj (engl. *layer normalization*) (Ba i drugi, 2016). Svrha ovog sloja je da ubrza vreme potrebno za treniranje modela.

3.2.5 *Feed forward sloj*

Nakon prolaska kroz *multi-head attention*, *residual* i *layer normalization* slojeve, aktivacije se dovode na ulaz *feed forward* sloja.

Svaka aktivacija prolazi kroz istu potpuno povezanu *feed forward* mrežu. Ovaj deo je moguće paralelizovati, jer aktivacije ne zavise jedna od druge tokom prolaska kroz *feed forward* mrežu.

3.3. *BERT (Bidirectional Encoder Representations from Transformers)*

Bidirectional Encoder Representations from Transformers, skraćeno *BERT*, je model predstavljen u radu (Devlin i drugi, 2018). *BERT-Base* model je izgrađen od 12 enkoder blokova originalnog transformera, naslaganih jedan na drugi (izlaz iz prethodnog enkodera je ulaz u naredni enkoder), dok je *BERT-Large* izgrađen od 24 enkoder blokova.

Treniran je nad *MLM (masked language modeling)* taskom, gde su neki tokeni u tekstu izostavljeni, a model treba da ih predvidi, uzimajući u obzir sve ostale tokene. Dodatno je treniran nad *NSP (next sentence prediction)* tasku, gde je zadatak da na osnovu dve prosleđene rečenice odredi da li je druga rečenica nastavak prve rečenice. Treniran je nad velikim korpusom knjiga i engleskom vikipedijom. Pokazano je da domen i dva taska nad kojim je treniran vrlo povoljno utiču na to da u svojim težinama sačuva reprezentaciju “svetskog znanja”. Tako istreniran model je moguće iskoristiti za rešavanje mnogih taskova koji zahtevaju razumevanje teksta (engl. *natural language understanding*).

4. SPECIFIKACIJA I IMPLEMENTACIJA REŠENJA

Ovo poglavlje posvećeno je korišćenim alatima (poglavlje 4.1), modelima neuronskih mreža (poglavlje 4.2), i skupovima podataka korišćenih za evaluaciju modela (poglavlje 4.3)

4.1. Korišćeni alati

Rezultati nad jednim delom skupova podataka su dobijeni upotrebom *SentEval* alata (Conneau i Kiela, 2018). Ovaj alat nudi jednostavnu evaluaciju *sentence embedding* modela u odnosu na više različitih skupova podataka. Većina korišćenih modela je preuzeta sa *sentence-transformers* repozitorijuma (<https://github.com/UKPLab/sentence-transformers>) kao pretrenirani modeli.

Svi eksperimenti su izvršeni na *Google Colaboratory* platformi (<https://colab.research.google.com>), a radna sveska (engl. *notebook*) u kojoj je urađen praktičan deo ovog rada i u kojoj je moguće u potpunosti reprodukovati dobijene rezultate, dostupna je na sledećem linku (<https://drive.google.com/file/d/1YHEUmCNd-aBPJxm9M2IgGc6t2Zh2ntRy/view?usp=sharing>).

4.2. Modeli neuronskih mreža

Kao pojedinačni modeli korišćeni su: više različitih modela sa *sentence-transformers* repozitorijuma (https://www.sbert.net/docs/pretrained_models.html#sentence-embedding-models) i *universal sentence encoder* (v5 *large* verzija, najnovija verzija u trenutku pisanja ovog rada) (<https://tfhub.dev/google/universal-sentence-encoder-large/5>). Kombinovani modeli predstavljaju kombinaciju više pojedinačnih modela.

4.2.1 Sentence transformers

Autori su pokazali da *BERT* mreža nije vremenski efikasna za pronalaženje semantički sličnih rečenica u velikoj kolekciji rečenica. Kao rešenje za taj problem nastaje *Sentence-BERT* (Reimers i Gurevych, 2019) model. Ovaj model predstavlja modifikaciju pretrenirane *BERT* (Devlin i drugi, 2018) mreže dodajući *pooling* sloj na izlaz *BERT* mreže kako bi se generisala vektorska reprezentacija rečenice fiksne dužine, nezavisno od dužine rečenice.

Tokom treniranja, da bi mreža naučila da proizvede semantički kvalitetne vektorske reprezentacije rečenica autori koriste *siamese* i *triplet* mreže (Schroff i drugi, 2015). Takođe, vrlo je bitno za koji problem će mreža biti prilagođena (Hill i drugi, 2016). Radovi (Conneau i drugi, 2017) i (Cer i drugi, 2018) dokazuju da prilagođavanje nad problemom zaključivanja u prirodnom jeziku (engl. *natural language inference*, skraćeno *NLI*) veoma pozitivno utiče na kvalitet. U problemu zaključivanja ulaz su dve rečenice, a izlaz da li je druga rečenica kontradikcija u odnosu na prvu, da li druga rečenica potvrđuje prvu rečenicu ili su te dve rečenice neutralne. Iz tog razloga *sentence transformers* su prilagođeni *Stanford Natural Language Inference (SNLI)* (Bowman i drugi, 2015) i *Multi-Genre NLI* (Williams i drugi, 2017) skupovima podataka. Modeli trenirani na ovaj način koji su korišćeni u ovom radu su: ***bert-base-nli-mean-tokens*** i ***nli-roberta-base-v2***.

Sledeća grupa modela je građena sa istom arhitekturom kao i *Sentence-BERT*, s' tim što umesto *BERT* kao osnove koristi *MPNet* (Song i drugi, 2020). *MPNet* je slična mreža kao i *BERT*, osim toga što je pretrenirana na drugačiji način. *MPNet* kombinuje pretrening *BERT* i *XLNet* (Yang i drugi, 2019) tako što koristi prednosti *BERT* pretreninga kako bi se nadomestile mane *XLNet* pretreninga, a takođe koristi prednosti *XLNet* pretreninga kako bi se nadomestile mane *BERT* pretreninga. Model treniran na ovaj način koji je korišćen u ovom radu je ***nli-mpnet-base-v2***.

Ostali modeli iste arhitekture koji su pored *NLI* prilagođeni dodatnim skupovima podataka su: ***sts-b-mpnet-base-v2*** - prilagođen *STS benchmark* (Cer i drugi, 2017) skupu podataka i ***paraphrase-mpnet-base-v2*** - prilagođen nad više skupova podataka za identifikaciju parafraze. Dodatan *sentence transformer* model koji je prilagođen *quora*

question pairs (<https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>) skupu podataka je *quora-distilbert-base*.

4.2.2 *Universal sentence encoder*

Autori u svom radu (Cer i drugi, 2018) predstavljaju dve arhitekture. Jedna je zasnovana na transformer arhitekturi, a druga na *DAN* (*deep averaging network*) (Iyyer i drugi, 2015). Transformer model postiže bolje performanse u odnosu na *DAN* model, ali poseduje kompleksniju arhitekturu koja zahteva više resursa. U ovom radu je korišćen transformer model.

Podaci za nenadgledano obučavanje transformer modela dobavljeni su sa različitih izvora, kao što su: *Wikipedia* (<https://www.wikipedia.org/>), vesti, sajtovi sa pitanjima i odgovorima i forumi za diskusije. *SNLI* skup podataka je potom korišćen za prilagođavanje.

Da bi se dobila vektorska reprezentacija rečenice fiksne dužine, model konvertuje sva slova u mala slova, rečenicu deli u tokene upotrebom *PTBTokenizer* (<https://nlp.stanford.edu/software/tokenizer.html>) i potom svaki token konvertuje u vektorsku reprezentaciju. Konačni izlaz iz modela se dobija tako što se saberu vektori aktivacija za sve tokene i potom se rezultujući vektor normalizuje da se ne bi stvarala razlika u semantici između dužih i kraćih rečenica.

4.2.3 **Kombinovani modeli neuronskih mreža**

Jedan kombinovani model čine 2 ili više pojedinačna modela. Kada svaki pojedinačni model generiše vektorsku reprezentaciju ulazne rečenice, tada se svi ti vektori konkatenuiraju u jedan zajednički vektor, što predstavlja izlaz kombinovanog modela.

Treba napomenuti da je isprobano više različitih kombinacija pojedinačnih modela u okviru ansambala, a da su navedeni samo oni koji će dati najinteresantnije rezultate. **Kombinovani 1** predstavlja kombinaciju *universal sentence encoder*, *paraphrase-mpnet* i *stsb-mpnet*. **Kombinovani 2** predstavlja kombinaciju *paraphrase-mpnet*,

quora-distilert i *stsb-mpnet*. **Kombinovani 3** predstavlja kombinaciju *nli-mpnet*, *paraphrase-mpnet* i *universal sentence encoder*.

4.3. Skupovi podataka

Za testiranje pojedinačnih modela i izbor za kombinovane modele korišćena je prva grupa skupova podataka (poglavlje 4.3.1), a za testiranje kombinovanih modela i upoređivanje sa najboljim modelima iz literature za konkretan skup podataka korišćena je druga grupa skupova podataka (poglavlje 4.3.2).

4.3.1 Prva grupa skupova podataka

Prvu grupu skupova podataka čine najčešće korišćeni skupovi podataka za identifikaciju semantičke sličnosti i detekciju parafraze. Iz tog razloga i iz razloga što je dosta pretreniranih modela ili prilagođeno ovim skupovima podataka ili su napravljeni imajući u vidu da će biti evaluirani nad ovim skupovima podataka, ovi skupovi podataka nisu korišćeni za krajnje upoređivanje performansi modela.

STS12-16 – Predstavlja grupu STS taskova po godinama, *STS 2012* (Agirre i drugi, 2012), *STS 2013* (Agirre i drugi, 2013), *STS 2014* (Agirre i drugi, 2014), *STS 2015* (Agirre i drugi, 2015) i *STS 2016* (Agirre i drugi, 2016). Svaki od ovih taskova je deo *SemEval* takmičenja (<https://semeval.github.io/>). Svake godine naučnici prezentuju nove anotirane skupove podataka koji imaju za cilj da pomognu u rešavanju različitih izazovnih problema na polju semantičke analize teksta. Više timova se takmiči i na kraju prezentuje svoje sisteme za rešavanje nekog od ovih problema koristeći nadgledano ili nenadgledano obučavanje.

STS benchmark (Cer i drugi, 2017) – Najčešće korišćen skup podataka za evaluaciju i upoređivanje performansi modela semantičke sličnosti rečenica. Sačinjen je od podskupa više različitih skupova podataka za semantičku sličnost sa *SemEval* takmičenja od 2012. do 2017. godine.

SICK-R (Marelli i drugi, 2014) - Izvor rečenica za ovaj skup podataka su opisi scena na slikama. Da bi se generisale rečenice sa sličnim značenjem autori su transformisali postojeće rečenice na različite načine: pojednostavljivanje kompleksnih rečeničnih konstrukcija u jednostavnije, zamena glagola ili imenica drugim glagolom ili imenicom sa sličnim značenjem, prebacivanje aktivnih rečenica u pasivne i obrnuto, izbacivanje ili dodavanje prideva, itd. Da bi se generisale rečenice sa različitim značenjem postojeće rečenice su transformisane na sledeće načine: ubacivanje ili izbacivanje negacija, zamena reči sa semantički različitim rečima, zamena redosleda reči u rečenici, itd.

MRPC (Dolan i Brockett, 2005) – Velika količina vesti na internetu je poslužila kao izvor podataka za ovaj skup podataka. Da bi se napravili parovi rečenica korišćene su metode bazirane na pravilima (engl. *rule based*) kao što su dužina rečenice, leksičko preklapanje, itd. Potom je *SVM* klasifikator treniran nad ručno definisanim obeležjima rečenica, što na kraju rezultuje u 5801 izabranih parova rečenica. Nakon toga dva anotatora su određivala labele za svaki par rečenica. Ukoliko se mišljenja razlikuju treći anotator je donosio finalnu odluku. Anotatorima nisu ponuđena jasna uputstva za razlikovanje rečenica koje su parafraze od onih koje nisu parafraze, pa su neki parovi rečenica potencijalno subjektivno ocenjeni.

4.3.2 Druga grupa skupova podataka

Drugu grupu skupova podataka čine uglavnom ređe korišćeni ili noviji skupovi podataka za identifikaciju semantičke sličnosti i detekciju parafraze. Iz tog razloga i iz razloga što svaki skup podataka predstavlja specifičan problem za rešavanje, ovi skupovi podataka će biti korišćeni za krajnje upoređivanje performansi modela. Primeri rečenica iz druge grupe skupova podataka su dati u tabeli 1, a kraći opisi u nastavku.

DSCS – Autori predstavljaju skup podataka (Chandrasekaran i Mago, 2020) čije su rečenice kompleksnije u odnosu na dotadašnje najčešće korišćene skupove podataka za semantičku sličnost. U radu je izvršena analiza čitljivosti rečenica. Dokazali su da su njihove rečenice značajno teže za čitanje i razumevanje u odnosu na rečenice iz *STS*

Benchmark i *SICK* skupova podataka. Svaka rečenica predstavlja definiciju jednog pojma iz domena računarskih nauka. Da bi dobavili rečenice sa sličnim značenjem, autori koriste tri različita izvora: *Wikipedia* (<https://www.wikipedia.org/>), *Simple English Wikipedia* (https://simple.wikipedia.org/wiki/Main_Page) i *the Merriam Webster Online dictionary* (<https://www.merriam-webster.com/>). Prilikom kreiranja parova rečenica, uparene su dve definicije iz različitih izvora. U anotiranju podataka je učestvovalo 5 studenata master studija sa usmerenja za računarske nauke i 10 radnika sa *Amazon Mechanical Turk (AMT)* platforme (<https://www.mturk.com/>) za koje je postojao zahtev da budu poznavaoici domena.

BIOSES – Skup podataka iz kojeg su izvučene rečenice za ovaj skup podataka (Soğancıoğlu i drugi, 2017) je predstavljen na konferenciji za analizu teksta, u okviru takmičenja za sumarizaciju biomedicinskih tekstova (<https://tac.nist.gov/2014/BiomedSumm/>). Autori su formirali parove rečenica sa pretpostavkom da dve rečenice koje citiraju isti rad, tj. sadrže istu referencu, su najverovatnije semantički slične. Korišćen je *pearson correlation coefficient* da bi se uporedila mera sličnosti između ocena 5 različitih anotatora. Vrednosti korelacije su između 0.9 i 0.96. Ocena sličnosti za svakog autora je data posebno, pa je za krajnju ocenu sličnosti jednog para rečenica korišćena aritmetička sredina ocena svih 5 autora.

Tabela 1. Primeri rečenica iz druge grupe skupova podataka

Skup podataka	Broj parova rečenica	Domen	Semantički sličan par rečenica	Semantički različit par rečenica
DSCS	50	Računarske nauke	A finite sequence of well-defined, computer-implementable instructions, typically to solve a class of problems or to perform a computation	A program can be a plan of how to do something
			A specific set of instructions or steps on how to complete a task.	A branch of science that deals with the theory of computation or the design of computers
BIOSSES	100	Biomedicina	It has recently been shown that Craf is essential for Kras G12D-induced NSCLC.	Here we show that both C/EBPa and NF1-A bind the region responsible for miR-223 upregulation upon RA treatment.
			It has recently become evident that Craf is essential for the onset of Kras-driven non-small cell lung cancer.	Isoleucine could not interact with ligand fragment 44, which contains amino group.
PAWS	8000	Opšte	Flights from Florida to New York	Flights from Florida to New York
			Flights from Florida to NYC	Flights from New York to Florida
PARADE	1358	Računarske nauke	the lowest level of code made up of 0s and 1s.	how the optimal solution to a linear programming problem changes as the problem data are modified.
			binary instructions used by the cpu.	how changes in the coefficients of a linear programming problem affect the optimal solution
PIT	972	Popularne teme na twitteru u tom trenutku	Aaaaaaaand stephen curry is on fire	Now lazy to watch Manchester united vs arsenal
			What a incredible performance from Stephen Curry	Early lead for Arsenal against Manchester United
TURL	10120	Razno	How an unverified but explosive dossier became a crisis for Donald Trump	Trump is manipulating the media the same way he manipulated the electorate
			How an Unverified Dossier Became a Crisis for @realDonaldTrump	Media must resist the urge to turn against one another . You're dealing with a master manipulator
OPUSPARCUS	1445	Razno	I'm disappointed in you .	I'm sorry , Doctor .
			You were a letdown .	You mustn't kill them , nurse .

PAWS - Par rečenica u ovom skupu podataka (Zhang i drugi, 2019) ima veliko leksičko preklapanje koje nisu parafraze. Rečenice koje nisu parafraze su uglavnom generisane zamenom reda reči u rečenici. Ovaj par rečenica: „*Flights from New York to Florida*“ i „*Flights from Florida to New York*“ ima potpuno poklapanje korišćenih reči u rečenici, a imaju suprotno značenje. Autori su pokazali da pretrenirani modeli bez prilagođavanja na ovom skupu podataka postižu preciznost (engl. *accuracy*) ispod 40%, dok nakon prilagođavanja dostižu do 85%. Postoje dve vrste ovog skupa podataka. U ovom radu će biti korišćen skup podataka za čije formiranje je korišćena *wikipedia*, poznatiji kao *PAWS wiki*.

PARADE – Za uspeh na ovom skupu podataka (He i drugi, 2020) je neophodno dobro poznavanje domena računarskih nauka. Parafraze imaju slabu leksičku i sintaksnu sličnost, ali po semantici su veoma slične ako se uzme u obzir poznavanje domena računarskih nauka. Parovi rečenica koje nisu parafraze imaju veliku leksičku i sintaksnu sličnost, a slabu semantičku sličnost. Pokazali su da *BERT* nakon prilagođavanja dostiže *f1* meru od samo 0,7. Anotatori koji nemaju domenskog znanja postižu još niže performanse. Rečenice su prikupljane sa kartica za podsećanje (engl. *flashcards*) sa raznih sajtova kao što je npr. (<https://quizlet.com/>). Na jednoj strani kartice je naziv nekog entiteta, a sa druge strane je definicija tog entiteta. Tokom prikupljanja podataka, autori su se vodili logikom da kartice sa istim entitetom verovatno imaju semantički slične definicije entiteta.

PIT – Ovaj skup podataka (Xu i drugi, 2015) je deo *SemEval* takmičenja iz 2015. godine. Podaci su dobijeni ekstrakcijom tvitova na teme koje su bile tada u trendu. Rečenice sa društvenih mreža često mogu biti gramatički neispravne, mogu sadržati reči iz slenga, netipične skraćenice, itd. što predstavlja poseban izazov za neuronske mreže pretrenirane na formalnijem tekstu u uslovima bez prilagođavanja. U anotiranju podataka su učestvovali 5 anotatora bez domenskog znanja i 5 anotatora sa domenskim znanjem.

TURL – Tvitovi koji sadrže isti link u sebi potencijalno govore o istoj temi koja se nalazi na deljenom linku i samim tim su potencijalno parafraze. Na ovaj način moguće je prikupiti oko 30,000 novih parafraza svakog meseca sa približno 70% preciznosti, što su autori dokazali u

radu (Lan i drugi, 2017). Ceo skup podataka sadrži 51,524 parova rečenica i predstavlja jedan od najvećih skupova podataka za identifikaciju parafraze.

OPUSPARCUS – Parovi rečenica su izvučeni iz *OpenSubtitles2016* korpusa (Lison i Tiedemann, 2016). Skup podataka sadrži titlove raznih filmova i serija. Ono što izdvaja ovaj skup podataka (Creutz, 2018) od ostalih je to što sadrži manje formalan tekst i što su rečenice deo konverzacije. Anotacija je izvršena na skali od 0 do 3 od strane dva anotatora. Ako se skor anotatora razlikuje za 1, onda se uzima manji skor kao finalni, a u slučaju da je razlika veća od 1, taj par rečenica se odbacuje.

5. METODOLOGIJA

Na slici 5 je prikazan tok određivanja semantičke sličnosti za jedan par rečenica. Rečenice prirodnog jezika bez bilo kakve prethodne obrade se dovode na ulaz dva jednaka, paralelna modela. Izlaz iz prvog i iz drugog modela su vektorske reprezentacije prve i druge rečenice, respektivno. Nakon toga se računa kosinusna sličnost između ta dva vektora. Kosinusna sličnost za dva vektora (A i B) se računa po formuli:

$$\text{similarity}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

Krajnji rezultat je vrednost između 0 i 1 koja reprezentuje meru semantičke sličnosti te dve rečenice. Vrednosti blizu 0 označavaju veoma malu semantičku sličnost, dok vrednosti blizu 1 označavaju veliku semantičku sličnost.

Da bi se odredile performanse modela za skupove podataka koji su pravljani u cilju evaluacije semantičke sličnosti korišćen je *pearson correlation coefficient* u odnosu na *ground truth* labela, koji se računa po formuli:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

gde je:

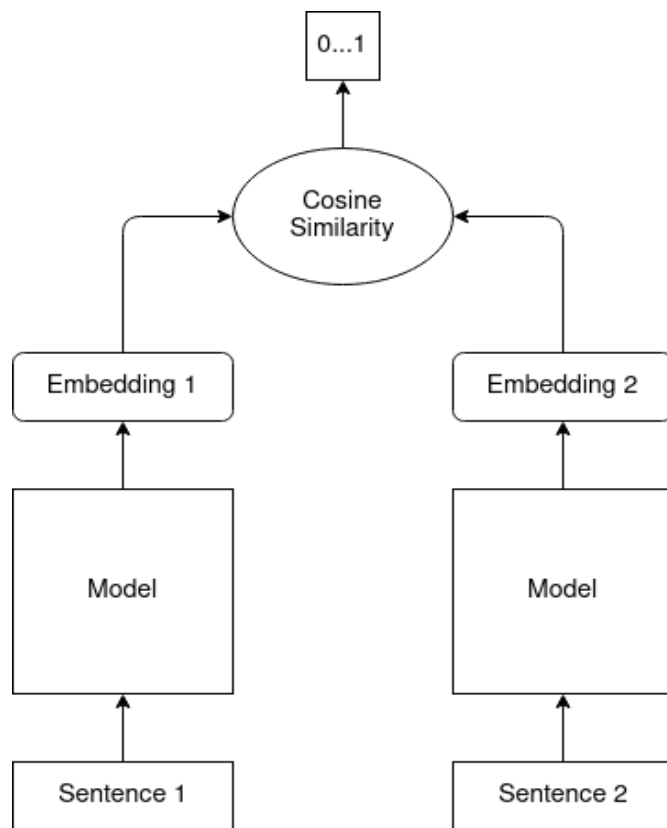
x_i - pojedinačna predikcija modela,

\bar{x} - srednja vrednost predikcija modela,

y_i - pojedinačna *ground truth* labela,

\bar{y} - srednja vrednost *ground truth* labela.

Za skupove podataka koji su pravljani u cilju evaluacije modela za identifikaciju parafraze korišćena je tačnost (engl. *accuracy*). Tačnost se računa tako što se broj tačno prediktovanih primeraka podeli sa ukupnim brojem primeraka. U slučaju detekcije parafraze, pošto je krajnji izlaz modela vrednost između 0 i 1, ta vrednost je zaokružena na bližu celobrojnu vrednost. Tada, ako je vrednost 1, par rečenica se klasifikuje kao parafraza, ako je vrednost 0 onda nije parafraza.



Slika 5. Šematski prikaz toka određivanja semantičke sličnosti dve rečenice (samostalno napravljena šema)

6. REZULTATI I DISKUSIJA

Svaki od pomenutih skupova podataka dolazi sa inicijalnom podelom na trening, validacioni i testni skup podataka (osim malih skupova podataka kao što su *DSCS* i *BIOSES* koji su u celini korišćeni kao testni skup podataka). Svi eksperimenti su izvršeni nad testnim skupom podataka. Dodatan razlog za tu odluku je što je svaki testni skup ručno anotiran od strane ljudi, dok su neki trening skupovi automatski anotirani.

6.1. Rezultati za prvu grupu skupova podataka

Rezultati pojedinačnih modela u odnosu na prvu grupu skupova podataka su dati u tabeli 2. Za *MRPC* skup podataka je korišćen *accuracy*, dok je za ostale skupove podataka korišćen *pearson correlation coefficient*. Svaki rezultat u tabeli je pomnožen sa 100.

Poređenja radi, u tabelu su dodati rezultati iz rada (Poerner i drugi, 2019). Autori kombinuju *Sentence-BERT*, *universal sentence encoder* i *ParaNMT* (Wieting i Gimpel, 2017) modele. Postižu najbolje rezultate u postojećoj literaturi za većinu skupova podataka iz prve grupe, u uslovima bez prilagođavanja ni jednom skupu podataka iz prve grupe.

Najbolji modeli bez prilagođavanja nekom od skupova podataka iz prve grupe su *paraphrase-mpnet* i *nli-mpnet*. Značajno lošije rezultate postiže *bert-base*, iz tog razloga neće biti evaluiran na drugoj grupi skupova podataka. Očekivano, najbolje rezultate postiže *stsb-mpnet*, koji je prilagođen *STSB* skupu podataka, a kao što je već navedeno u sklopu tog skupa podataka se nalaze neki skupovi podataka iz *STS12-16*.

Tabela 2. Rezultati za prvu grupu skupova podataka

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	MRPC	Prosek
Pretrenirani modeli									
USE(v5-large)	70.94	70.47	75.95	82.86	81.29	81.49	87.57	68.35	77.37
bert-base-nli-mean-tokens	64.61	67.54	73.22	74.34	70.13	74.15	84.24	67.48	71.96
nli-roberta-base-v2	69.76	79.05	81.37	82.85	80.98	84.71	86.65	68.41	79.22
nli-mpnet-base-v2	71.34	79.97	83.59	85.05	82.01	85.73	87.96	68.35	80.50
paraphrase-mpnet-base-v2	72.90	80.47	82.50	83.02	83.00	86.25	88.34	69.45	80.74
quora-distilbert-base	71.31	75.99	84.18	79.98	76.44	79.72	85.18	70.55	77.92
(Poerner i drugi, 2019)	72.80	69.60	81.70	84.20	81.30	83.90	/	/	78.92
Pretrenirani modeli + prilagođeni STSb									
sts-b-mpnet-base-v2	75.93	84.23	93.01	88.07	88.07	88.16	88.07	73.04	84.82

Za evaluaciju nad drugom grupom skupova podataka biće uglavnom uzeti modeli koji su pokazali najbolje performanse na prvoj grupi skupova podataka. Iz razloga što postiže slične performanse kao i drugi bolji modeli, *nli-roberta* neće biti evaluirana nad drugom grupom. Iako *universal sentence encoder* postiže nešto lošije rezultate u odnosu na većinu evaluiranih modela, izabran je da bude evaluiran nad drugom grupom skupova podataka iz razloga što poseduje drugačiju arhitekturu u odnosu na ostale modele, što može potencijalno pozitivno uticati na neke skupove podataka iz druge grupe. Takođe, pored lošijih performansi, *quora-distilbert* model je izabran za dalju evaluaciju iz razloga što je ovaj model prilagođen *quora* skupu podataka, što može potencijalno pozitivno uticati na performanse u okviru kombinovanog modela.

6.2. Rezultati i diskusija za drugu grupu skupova podataka

Za različite skupove podataka su korišćene različite metrike kako bi se rezultati mogli uporediti sa drugim rezultatima iz literature. Za *DSCS*, *BIOSSES* i *PIT* je korišćen *pearson correlation coefficient*, za *TURL* *average precision* a za sve ostale *accuracy*.

Rezultati izabranih pojedinačnih modela i modela ansambala u odnosu na drugu grupu skupova podataka su dati u tabeli 3. Svaki rezultat u tabeli je pomnožen sa 100.

Može se primetiti da se prosečni rezultati pojedinačnih modela ne razlikuju mnogo, manje od 2 procenta, ako se izuzme *quora-distilbert* model koji se značajno lošije pokazao u odnosu na ostale pojedinačne modele na svim skupovima podataka. Jedan od mogućih razloga zašto *quora-distilbert* model postiže lošije rezultate je to što su u *quora* skupu podataka parovi upitnih rečenica, dok su u svim ostalim skupovima podataka iz druge grupe parovi izjavnih rečenica. Još jedna od hipoteza je da je *quora* skup podataka suviše specifičan, tj da modeli obučeni na njemu ne mogu dovoljno dobro da generalizuju i da budu uspešni na drugim skupovima podataka.

Ne postoji model koji za sve skupove podataka postiže najbolji rezultat, za različite skupove podataka različiti modeli postižu najbolji rezultat. Pojedinačni model koji je postigao najbolji prosečni rezultat je *paraphrase-mpnet*. Iako ovaj model nije najčešće najbolji model za pojedinačni skup podataka, on za svaki skup podataka postiže približno najbolje rezultate.

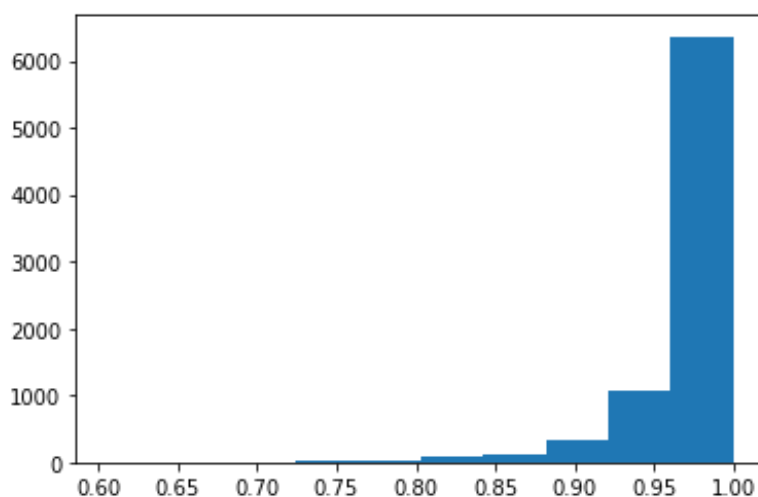
Posmatrajući *universal sentence encoder* model, on postiže najbolje rezultate na *TURL* i skoro najbolje rezultate na *PIT* skupu podataka. Pošto je *universal sentence encoder* pretreniran nad tekstom sa raznih foruma i sajtova, to mu daje prednost nad skupovima podataka koji su nastali od tvitova i sadrže dosta gramatički neispravnih reči i reči iz slenga.

Tabela 3. Rezultati za drugu grupu skupova podataka

Model	DSCS	BIOSSES	PAWS	PARADE	PIT	TURL	OPUSPARCUS	Prosek
USE (v5-large)	75.74	69.15	44.20	72.37	65.99	78.50	82.63	69.80
nli-mpnet-base-v2	84.00	71.34	44.20	60.13	59.09	76.87	86.71	68.91
paraphrase-mpnet-base-v2	83.39	82.32	44.20	63.52	66.42	76.33	85.47	71.66
stsb-mpnet-base-v2	82.56	81.87	44.21	66.62	58.91	75.82	85.40	70.77
quora-distilbert-base	68.04	67.91	44.24	57.33	47.40	70.31	84.71	62.85
Kombinovani 1	86.04	84.54	44.20	65.81	65.72	77.21	86.16	72.81
Kombinovani 2	75.80	73.23	44.20	58.81	55.82	74.03	86.30	66.88
Kombinovani 3	86.27	80.59	44.20	61.97	66.28	77.76	86.78	71.98
sota iz literature	73.67	93.80	91.90	75.30	75.8	81.60	89.20	83.04

Najbolji kombinovani model je sastavljen od 3 najuspešnija pojedinačna modela i postiže za oko 1 procenat bolji prosečan rezultat od najboljeg pojedinačnog modela. Rezultati za Kombinovani 2 pokazuju da *quora-distilbert* model kao deo kombinovanog modela loše utiče na celokupne performanse. Kombinovani model koji je izgrađen od svih pojedinačnih modela evaluiranih nad drugom grupom skupova podataka, a koji nije naveden u tabeli 3 postiže prosečni rezultat od 68,03 što je lošiji rezultat u odnosu na gotovo sve pojedinačne modele. Najverovatnije je uzrok tome prisustvo *quora-distilbert* modela u ansamblu, kao i za Kombinovani 2 model.

Ni jedan model ne radi dobro na *PAWS* skupu podataka, čak svi postižu skoro identične performanse. Na slici 6 je prikazan histogram predikcija za *paraphrase-mpnet* modela. Iz priloženog histograma vidi se da model apsolutno sve parove rečenica klasifikuje kao parafraze. Ovo je u skladu sa rezultatima dobijenim u (Zhang i drugi, 2019) gde su autori pokazali da modeli koji nisu prilagođeni ovom skupu podataka postižu loše performanse. Razlog tome je što postojeći skupovi podataka sadrže veoma mali broj parova rečenica koje sadrže veliko leksičko preklapanje a nisu parafraze. Većina posmatranih modela je pretrenirano na *SNLI* skupu podataka, dok su autori u (Dasgupta i drugi, 2018) pokazali kako većina kontradikcija u *SNLI* skupu podataka ima veoma malo leksičko preklapanje te modeli pretrenirani na ovom skupu loše klasifikuju kontradikcije sa velikim leksičkim preklapanjem.



Slika 6. Histogram predikcija *paraphrase-mpnet* modela za *PAWS* skup podataka

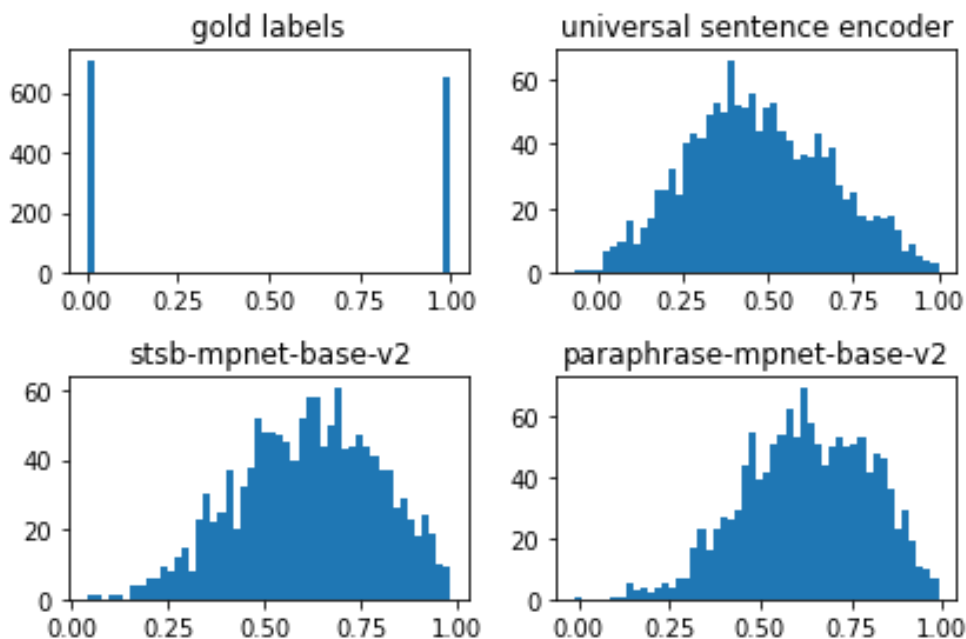
Poslednji red u tabeli 3 predstavlja najbolje rezultate za svaki skup podataka iz literature. Slede reference na radove odakle su uzeti rezultati: *DSCS* (Chandrasekaran i Mago, 2020), *BIOSSES* (Kanakarajan i drugi, 2021), *PAWS* (Zhang i drugi, 2019), *PARADE* (He i drugi, 2020), *PIT* (Wang i drugi, 2021), *TURL* (Wang i drugi, 2021) i *OPUSPARCUS* (Sjöblom i drugi, 2018). Najbolji rezultat iz literature za *DSCS* skup podataka je dobijen tako što je Albert model (Lan i drugi, 2019) iskorišćen kao bazni model umesto *Bert* modela u *Sentence-Bert* arhitekturi koji je potom prilagođen *STS Benchmark* skupu podataka. Za *BIOSSES* skup podataka najbolji rezultat je proizveo *ELECTRA* model (Clark i drugi, 2020) koji je pretreniran nad velikom količinom teksta iz domena medicine. Za sve skupove podataka osim *DSCS* i *BIOSSES* najbolji rezultati iz literature su dobijeni treniranjem ili prilagođavanjem nad trening delom skupa podataka. Kao što je očekivano, bolje rezultate postižu modeli koji su ili trenirani nad ili prilagođeni trening delu skupa podataka.

Da bi se identifikovala razlika između modela koji nisu prilagođeni ni jednom skupu podataka iz druge grupe i onih koji jesu, upoređen je prosečan rezultat za najbolji kombinovani model sa prosekom najboljih rezultata iz literature, posmatrajući *PARADE*, *PIT*, *TURL* i *OPUSPARCUS* skupove podataka. Tada je prosečan rezultat za kombinovani 73,73 a za modele iz literature 80,48 što je približno 7 procenata razlike.

6.2.1 Analiza predikcija modela za *PARADE* skup podataka

U tabeli 3 se mogu identifikovati rezultati koji se ne mogu lako objasniti, tj. bez analize predikcija modela nije moguće zaključiti zašto su neki modeli uspešniji ili lošiji za neki skup podataka u odnosu na druge modele. Jedan od skupova podataka na kome se performanse nekih modela značajno razlikuju od drugih je *PARADE*.

Labele u *PARADE* skupu podataka su balansirane, 650 parafraza i 707 neparafraza. Histogrami na slici 7 pokazuju da distribucija predikcija za *universal-sentence-encoder* gotovo odgovara distribuciji labela uz blagu naklonost ka neparafrazama, dok ostali modeli imaju sklonost da većinu parova rečenica klasifikuju kao parafraze.



Slika 7. Histogrami predikcija izabranih modela za PARADE skup podataka

Posmatrajući matrice konfuzije (engl. *confusion matrix*) na slici 8, uočava se da *universal sentence encoder* odstupa od pravilnosti koju čine svi ostali modeli. *Universal sentence encoder* ima najveći broj *false negative* a najmanji broj *false positive* predikcija u odnosu na sve ostale modele, dok svi ostali modeli imaju vrlo veliki broj *false positive* predikcija a dosta mali broj *false negative* predikcija. Nešto bolje rezultate među ostalim modelima postiže *stsb-mpnet-base-v2* model koji ima nešto manje *false positive* predikcija od ostalih. Od kombinovanih modela Kombinovani 1 postiže najbolje rezultate, najverovatnije iz razloga što su i *universal sentence encoder* i *stsb-mpnet-base-v2* deo ovog modela. Na slici 8 su prikazani histogrami predikcija za *universal-sentence-encoder*, *stsb-mpnet-base-v2* i *paraphrase-mpnet-base-v2* modele.

Analizirajući primere koji su označeni kao parafraze a koje su svi modeli klasifikovali kao neparafraze uočene su sledeće pravilnosti. Primeri rečenica: “the lowest level of code made up of 0s and 1s.”, “binary instructions used by the cpu.”, “low level programming language which programs for a specific cpu. uses binary.” i “original low level language, rows of switches flipped on or off by operator” su međusobno sve parafraze, međutim svi modeli ih klasifikuju kao

neparafraze. Ovo su primeri gde je neophodno domensko poznavanje računarskih nauka kako bi se pravilno klasifikovali.

		PREDICTED					
		T	F	universal sentence encoder		nli-mpnet-base-v2	
ACTUAL	T	True positive	False negative	431	219	615	35
	F	False positive	True negative	156	551	506	201
		paraphrase-mpnet-base-v2		sts-b-mpnet-base-v2		quora-distilbert-base	
	T	594	56	589	61	612	38
	F	439	268	392	315	541	166
		Kombinovani 1		Kombinovani 2		Kombinovani 3	
	T	593	57	614	36	607	43
	F	407	300	523	184	473	234

Slika 8. Matrice konfuzije za sve modele na PARADE skupu podataka. Pozitivnu klasu čine parafraze a negativnu neparafraze.

Primeri koji su označeni kao parafraze, koje *universal sentence encoder* klasifikuje kao neparafraze a svi modeli su ih klasifikovali kao parafraze su dati u nastavku. Rečenica “*people who use the applications and databases*” je parafraza za “*use data for queries and reports. some even update the database content.*” i “*the people whose jobs require access to the database for querying, updating, and generating reports.*”. Moguće je da *universal sentence encoder* “uočava” male razlike u značenju ovih rečenica, jer se u prvoj rečenici pominje *applications*, a za onog ko nije poznavalac domena nije jasno da se tu misli na aplikacije povezane sa radom nad bazom podataka. Druga mogućnost zašto su ostali modeli bolji na polju *false negative* je to što oni imaju sistematsko odstupanje (engl. *bias*), tj. većinu rečenica iz ovog skupa podataka

klasifikuju kao parafraze, bez stvarnog “razumevanja” šta je parafraza a šta ne.

Primeri koji su označeni kao neparafraze, koje *universal sentence encoder* klasifikuje kao neparafraze a svi modeli su ih klasifikovali kao parafraze su dati u nastavku. Rečenica “*rarely written in machine code (binary) as it is harder to understand*” nije parafraza za “*low level programming language which programs for a specific cpu. uses binary.*” Ovde se jasno vidi da te dve rečenice ne znače isto, međutim imaju neki nivo semantičke sličnosti, jer pričaju o istoj stvari. Rečenica “*small computers that the average person can use*” nije parafraza za “*computers typically used by one user at home or in office. generally used for general computer tasks.*” i “*a computing device designed to be used by one person at a time.*”. U ovom slučaju je teže razumeti zašto ovi parovi rečenica nisu parafraze. Moguće je da su ovo primeri dodeljivanja pogrešne labele a takođe se može doneti zaključak da je problem semantičke sličnosti dosta zavisao od anotatora, tj. njihovog subjektivnog osećaja.

7. ZAKLJUČAK

Motivacija ovog rada je bila da se istraži da li postoji *sentence embedding* model koji radi dobro za bilo koji domen i bilo koju vrstu teksta na engleskom jeziku za problem semantičke sličnosti rečenica.

Upoređene su performanse pojedinačnih *state-of-the-art sentence embedding* modela, kao i njihovih ansambala, sa modelima iz literature. *Sentence embedding* modeli su ili prilagođeni nekom drugom problemu ili su prilagođeni nekom skupu podataka za semantičku sličnost koji se nije koristio za krajnju evaluaciju modela. Izlaz kombinovanog modela predstavlja konkatenirane vektorske reprezentacije rečenica, dobijenih od više pojedinačnih modela. Modeli iz literature su trenirani nad tekstem iz istog domena, direktno trenirani ili prilagođeni skupovima podataka korišćenih za krajnje upoređivanje rezultata.

Skupovi podataka su podeljeni u dve grupe. Prvu grupu čine skupovi podataka koji su često korišćeni u literaturi. Ova grupa je iskorišćena kako bi se isprobali svi pojedinačni *sentence embedding* modeli i kako bi se izabrali kandidati za evaluaciju na drugoj grupi. Drugu grupu čine skupovi podataka koji su ređe korišćeni u literaturi, a koji su pravljeni u cilju evaluacije modela za neki specifičan domen ili specifične strukture rečenica. Iz ovih razloga druga grupa je iskorišćena za krajnju evaluaciju, gde su upoređene performanse pojedinačnih modela sa kombinovanim i rezultatima iz literature za konkretne skupove podataka.

Da bi se odredila mera semantičke sličnosti dve rečenice, svaka rečenica je posebno, bez prethodne obrade, dovedena na ulaz modela, a izlaz je vektorska reprezentacija te rečenice. Nakon toga se primenjuje kosinusna sličnost između ta dva vektora. Rezultat je vrednost između 0 i 1. Vrednosti bliže 1 znače da su rečenice semantički slične, a vrednosti bliže 0 znače da nisu.

Jednostavnom konkatenacijom vektora rečenica najbolji kombinovani model postiže za oko 1 procenat bolji skor u odnosu na najbolji pojedinačni model. Da bi se dalje unapredio ovaj pristup trebalo bi isprobati različite metode kreiranja ansambala. Poredeći srednju vrednost rezultata nad podskupom druge grupe skupova podataka, modeli iz literature koji su prilagođeni svakom skupu podataka posebno

su postigli za 7 procenata bolji rezultat u odnosu na najbolji kombinovani model. Ovo nije tako velika razlika, što znači da se oblast bliži cilju da kreira jedan model koji će raditi dobro ne vezano za domen i vrstu teksta.

Analiziranjem predikcija modela na *PARADE* skupu podataka uočavaju se primeri iz skupa podataka za koje je se predikcije većina modela razlikuju od ciljne labele. U nekim slučajevima je moguća greška anotatora dok je u ostalim slučajevima diskutabilno da li je labela odgovarajuća. Iz toga se izvodi zaključak da je interpretacija semantičke sličnosti često subjektivna i da ne postoji generalna skala nivoa semantičke sličnosti koja bi se mogla primeniti za sve skupove podataka.

Kao nastavak ovog rada, bilo bi interesantno ispitati da li će prilagođavanje najboljeg kombinovanog modela jednom od skupova podataka iz druge grupe drastično uticati na poboljšanje performansi na svim ostalim skupovima podataka iz druge grupe.

8. LITERATURA

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., ... & Wiebe, J. (2014). Semeval-2014 task 10: Multilingual semantic textual similarity. In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014) (pp. 81-91).

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., ... & Wiebe, J. (2015). Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015) (pp. 252-263).

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez Agirre, A., Mihalcea, R., ... & Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511. ACL (Association for Computational Linguistics).

Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) (pp. 385-393).

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., & Guo, W. (2013). * SEM 2013 shared task: Semantic textual similarity. In Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity (pp. 32-43).

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 169-174).

Chandrasekaran, D., & Mago, V. (2020). Domain Specific Complex Sentence (DSCS) Semantic Similarity Dataset. arXiv preprint arXiv:2010.12637.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Conneau, A., & Kiela, D. (2018). Senteval: An evaluation toolkit for universal sentence representations. arXiv preprint arXiv:1803.05449.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.

Cordeiro, J. P., Dias, G., & Brazdil, P. (2007). Unsupervised Learning of Paraphrases. *Research in Computer Science*. National Polytechnic Institute, Mexico. ISSN, 4069.

Creutz, M. (2018). Open subtitles paraphrase corpus for six languages. *arXiv preprint arXiv:1809.06142*.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., & Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Deshmukh, A. A., & Sethi, U. (2020). IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles. *arXiv preprint arXiv:2007.12603*.

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*.

Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Grootendorst, M. (2020). BERTopic: leveraging BERT and c-TF-IDF to create easily interpretable topics, URL <https://doi.org/10.5281/zenodo.4381785>.

He, Y., Wang, Z., Zhang, Y., Huang, R., & Caverlee, J. (2020). PARADE: A new dataset for paraphrase identification requiring computer science domain knowledge. *arXiv preprint arXiv:2010.03725*.

Hill, F., Cho, K., & Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

- Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers) (pp. 1681-1691).
- Kanakarajan, K., Kundumani, B., & Sankarasubbu, M. (2021). BioELECTRA: Pretrained Biomedical text Encoder using Discriminators. In Proceedings of the 20th Workshop on Biomedical Language Processing (pp. 143-154).
- Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294-3302).
- Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.
- Lan, W., Qiu, S., He, H., & Xu, W. (2017). A continuously growing dataset of sentential paraphrases. arXiv preprint arXiv:1708.00391.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady (Vol. 10, No. 8, pp. 707-710).
- Linnainmaa, S. (1970). The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's Thesis (in Finnish), Univ. Helsinki, 6-7.
- Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.
- Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., & Zamparelli, R. (2014). A SICK cure for the evaluation of

compositional distributional semantic models. In *Lrec* (pp. 216-223).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Poerner, N., Waltinger, U., & Schütze, H. (2019). Sentence meta-embeddings for unsupervised semantic textual similarity. *arXiv preprint arXiv:1911.03700*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Ranasinghe, T., Orasan, C., & Mitkov, R. (2019). Enhancing unsupervised sentence similarity methods with deep contextualised word representations. *RANLP*.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sjöblom, E., Creutz, M., & Aulamo, M. (2018). Paraphrase detection on noisy subtitles in six languages. arXiv preprint arXiv:1809.07978.

Soğancıoğlu, G., Öztürk, H., & Özgür, A. (2017). BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14), i49-i58.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. arXiv preprint arXiv:2004.09297.

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).

Wang, K., Reimers, N., & Gurevych, I. (2021). TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning. arXiv preprint arXiv:2104.06979.

Wieting, J., & Gimpel, K. (2017). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. arXiv preprint arXiv:1711.05732.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Xu, W., Callison-Burch, C., & Dolan, W. B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 1-11).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhang, Y., Baldridge, J., & He, L. (2019). PAWS: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

BIOGRAFIJA

Marko Radović je rođen 02.07.1996. godine u Novom Sadu. Osnovnu školu “Laza Kostić” u Kovilju završio je 2011. godine kao nosilac diplome „Vuk Karadžić“. Prirodno-matematički smer gimnazije “Isidora Sekulić” završava 2015. godine, takođe kao nosilac diplome „Vuk Karadžić“. Iste godine upisao je Fakultet tehničkih nauka, smer softversko inženjerstvo i informacione tehnologije. Tokom studija bio je član “EESTEC LC Novi Sad” studentske organizacije i učestvovao u organizaciji događaja u “Startit” centru u Novom Sadu. U julu 2017. godine obavio je jednomesečnu stručnu praksu u kompaniji “Execom”. U februaru 2018. postaje stipendista iste kompanije, radeći *part-time*. U februaru 2021. postaje *full-time* zaposlen u kompaniji “HTEC”.

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj, RBR:	
Identifikacioni broj, IBR:	
Tip dokumentacije, TD:	monografska publikacija
Tip zapisa, TZ:	tekstualni štampani dokument
Vrsta rada, VR:	diplomski rad
Autor, AU:	Marko Radović
Mentor, MN:	Dr Aleksandar Kovačević, vanredni profesor
Naslov rada, NR:	Poređenje modela za identifikaciju semantičke sličnosti rečenica
Jezik publikacije, JP:	srpski
Jezik izvoda, Ji:	srpski / engleski
Zemlja publikovanja, ZP:	Srbija
Uže geografsko područje, UGP:	Vojvodina
Godina, GO:	2021
Izdavač, IZ:	autorski reprint
Mesto i adresa, MA:	Novi Sad, Fakultet tehničkih nauka, Trg Dositeja Obradovića 6
Fizički opis rada, FO:	(broj poglavlja 8/ stranica 52/ slika 8/ grafikona 0/ tabela 3/ referenci 54/ priloga 0)
Naučna oblast, NO:	Softversko inženjerstvo i informacione tehnologije
Naučna disciplina, ND:	Mašinsko učenje
Predmetna odrednica / ključne reči, PO:	mašinsko učenje, transformer neuronske mreže
UDK	
Čuva se, ČU:	Biblioteka Fakulteta tehničkih nauka, Trg Dositeja Obradovića 6, Novi Sad
Važna napomena, VN:	
Izvod, IZ:	U ovom radu je evaluirano više pretreniranih <i>sentence embedding</i> modela u odnosu na više skupova podataka za semantičku sličnost i identifikaciju parafraze u cilju upoređivanja performansi modela u odnosu na skupove podataka.
Datum prihvatanja teme, DP:	
Datum odbrane, DO:	
Članovi komisije, KO:	
predsednik	Dr Jelena Slivka, vanredni profesor
član	Dr Nikola Luburić, docent
mentor	Dr Aleksandar Kovačević, vanredni profesor
Potpis mentora	

KEY WORDS DOCUMENTATION

Accession number, ANO :	
Identification number, INO :	
Document type, DT :	monographic publication
Type of record, TR :	textual material
Contents code, CC :	bachelor thesis
Author, AU :	Marko Radović
Mentor, MN :	Aleksandar Kovačević, associate professor, PhD
Title, TI :	Comparison of models for identifying the semantic similarity of sentences
Language of text, LT :	Serbian
Language of abstract, LA :	Serbian / English
Country of publication, CP :	Serbia
Locality of publication, LP :	Vojvodina
Publication year, PY :	2021
Publisher, PB :	author's reprint
Publication place, PP :	Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6
Physical description, PD :	(number of chapters 8/ pages 52/ figures 8/ graphs 0/ tables 3/ references 54/ attachment 0)
Scientific field, SF :	Software Engineering and Information Technologies
Scientific discipline, SD :	Machine Learning
Subject / Keywords, S/KW :	machine learning, transformer neural networks
UDC	
Holding data, HD :	Library of the Faculty of Technical Sciences, Trg Dositeja Obradovića 6, Novi Sad
Note, N :	
Abstract, AB :	In this paper, multiple pretrained sentence embedding models are evaluated against multiple datasets for semantic textual similarity and paraphrase identification in order to compare their performance per specific dataset.
Accepted by sci. Board on, ASB :	
Defended on, DE :	
Defense board, DB :	
president	Jelena Slivka, associate professor, PhD
member	Nikola Luburić, assistant professor, PhD
mentor	Aleksandar Kovačević, associate professor, PhD
Mentor's signature	