Reprezentativno učenje u obradi prirodnog jezika

Pribanić, Ana

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2020./2021.

Ana Pribanić

Reprezentativno učenje u obradi prirodnog jezika

Završni rad

Mentorica: prof.dr.sc. Nives Mikelić Preradović

Zagreb, srpanj 2021.

Izjava o akademskoj čestitosti

Izjavljujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

(potpis)

Sadržaj

1. Uvod

Dosadašnji napredak informatičke tehnologije omogućio nam je lako i efikasno pretraživanje informacija, automatizirane glasovne asistente, automatsko ispravljanje pogrešaka u tekstu, filtriranje elektroničke pošte, automatsko generiranje predložaka tekstova, prevođenje tekstova s jednog jezika na drugi i mnoge druge slične usluge koje zahtijevaju interakciju između računala i ljudskog – prirodnog – jezika. Za njihov razvoj i uspješnu izvedbu posebno je zaslužan značajan napredak ostvaren u području obrade prirodnog jezika. Budući da se ljudi sve više oslanjaju na računalne sustave za komunikaciju i izvršavanje zadataka, javlja se potreba za usavršavanjem za to zaslužnih tehnologija. Veliki napredak u tom kontekstu je ostvario Google s razvojem modela transformera koji su implementirani u Google prevoditelj i Google-ov sustav za pretraživanje. Korištenjem novih tehnika i pristupa, ovi modeli su se pokazali vrlo uspješnima u razumijevanju i generiranju tekstova na prirodnom jeziku. U prvom dijelu ovoga rada dan je teorijski pregled primjene i područja obrade prirodnog jezika. Objašnjena je važnost uloge reprezentativnog učenja unutar tog područja te je dana podjela na temelju vrsta reprezentacija. U drugom dijelu rada je demonstrirana vizualizacija skupa podataka pomoću modela transformera na temelju čega su istaknute poteškoće s kojima su se modeli susreli. Na kraju je napravljena usporedba različitih modela transformera s obzirom na njihovu učinkovitost pri vizualizaciji korištenog skupa podataka.

2. Obrada prirodnog jezika

Za razliku od umjetnih i formalnih jezika, prirodni jezici su se razvili bez svjesnog planiranja ili predumišljaja. Obrada prirodnog jezika (engl. *Natural Language Processing - NLP*) stoga pripada području računalnih znanosti, ali i području društvenih znanosti, polju informacijskih znanosti, gdje postoji grana obrada prirodnog jezika, leksikografija i enciklopedika. Računalna je lingvistika više usmjerena na aspekte jezika, dok OPJ naglašava njegovu uporabu u strojnom učenju i tehnikama dubokog učenja za izvršavanje zadataka (Šuman, 2021.). Cilj OPJ-a je osposobiti računala da razumiju tekst i izgovorene riječi na sličan način na koji to mogu i ljudi. To je izuzetno zahtjevan i kompleksan proces upravo zbog naravi prirodnih jezika koji obuhvaćaju bogate rječnike, višeznačnost riječi, dijalekte, različite naglaske, žargone, sarkazam, ironiju, itd. Iako OPJ još uvijek nailazi na prepreke savladavanja ljudskog jezika, zabilježen je veliki napredak posljednjih godina u ovom području.

Razvoj OPJ-a možemo pratiti od začetka ideje strojnog prevođenja koja se pojavila tijekom drugog svjetskog rata (Liddy, 2001). Tada je primarna ideja bila korištenje računala za prevođenje teksta s jednog jezika na drugi. U početku se rezultati nisu pokazali uspješnima upravo iz razloga što nije uzeta u obzir leksička višeznačnost svojstvena prirodnom jeziku, ali i ograničenja tadašnje tehnologije. Napredak ostvaren u području lingvistike, preciznije Chomskyjeva teorija generativne gramatike, uvelike je doprinijela boljem shvaćanju funkcioniranja strojnog prevođenja te potaknula druge primjene u OPJ-u poput prepoznavanja govora (Liddy, 2001). Krajem 1960-ih i početkom 1970-ih uz niz teoretskih pristupa kao reakcije na Chomskyjevu teoriju, osmišljeni su i razni prototipi sustava koji su demonstrirali djelotvornost pojedinih lingvističkih principa (Liddy, 2001). ELIZA je jedan od prvih programa OPJ-a koji je bio predstavljen kao terapeut. Funkcionirao je tako da bi analizirao ključne riječi u ulaznim rečenicama korisnika i na temelju njih reproducirao odgovor uz pomoć prethodno implementiranog scenarija (Weizenbaum, 1966). Iako program nije imao istinsko shvaćanje značenja ulaznih riječi i rečenica, dokazao je da je komunikacija između računala i čovjeka ostvariva. Današnji glasovni asistenti su puno sofisticiraniji te mogu uz određenu naredbu ispunjavati svakodnevne zadatke uključujući obavljanje poziva, unošenje podsjetnika, odgovaranje na pitanja i slično. U posljednjih deset godina OPJ je ostvario obećavajući napredak koji možemo pripisati dostupnosti sve većem broju podataka u digitalnom obliku, ubrzanom radu računala i veličini memorije te pojavi interneta.

2.1. Primjena

OPJ obuhvaća različite aplikacije za olakšavanje interakcije između ljudskog jezika i računala koje su u svakodnevnoj uporabi. Njegova primjena važna je za različite tvrtke koje posjeduju velike količine nestrukturiranog teksta, poput elektroničke pošte, razgovora na društvenim mrežama, anketa, knjiga, internetskih stranica, zdravstvenih kartona, itd. Alati za obradu prirodnog jezika mogu pomoći tvrtkama da analiziraju podatke i dobiju uvide u korisničko iskustvo te automatiziraju dugotrajne procese. Širok je spektar primjene OPJ-a, od jednostavne analize teksta do njegove kompleksnije i zahtjevnije obrade. Neke od njegovih najčešćih uporaba su (Liddy, 2001):

- Pretraživanje informacija (engl. *Information Retrieval - IR*) – podrazumijeva pronalaženje dokumenata nestrukturiranog teksta u velikim kolekcijama koji odgovaraju informacijskoj potrebi korisnika (Ahmad & Dang, 2014) . Google pretraživač najpoznatiji je sustav za pretraživanje informacija koji prepoznaje dokumente na temelju dane riječi ili rečenice.

- Izdvajanje informacija (engl. *Information Extraction* – IE) – zadatak automatskog izdvajanja strukturiranih podataka iz nestrukturiranih i/ili polu-strukturiranih strojno čitljivih dokumenata i drugih elektronički zastupljenih izvora (Ahmad & Dang, 2014). Glavna razlika između IR i IE je ta što su relevantne informacije kod IE-a unaprijed definirane, dok IR pokušava pronaći dokumente koji odgovaraju na korisnikovu potrebu, a za koje korisnik nije svjestan.

- Analiza sentimenta (engl. *Sentiment Analysis*) – koristi OPJ, analizu teksta i računalne tehnike za automatsko izdvajanje ili klasifikaciju sentimenta iz teksta (Hussein. 2018). Koristi se u razne svrhe kao što su prikupljanje povratnih informacija potrošača za određene proizvode, praćenje i mjerenje sentimenta na društvenim mrežama, istraživanje tržišta itd.

- Odgovaranje na pitanja (engl. *Question-Answering*) – sustavi koji automatski odgovaraju na pitanja postavljena prirodnim jezikom koristeći unaprijed strukturiranu bazu podataka ili zbirku dokumenata na prirodnom jeziku. Omogućuju postavljanje

pitanja i pružanje odgovora pomoću upita na prirodnom jeziku te se mogu smatrati naprednim oblikom IR-a (Calijorne Soares & Parreiras, 2020).

- Strojno prevođenje (engl. *Machine Translation - MT*) – skup alata koji omogućavaju korisnicima unos teksta na jednom jeziku te generiranje prijevoda na drugi ciljani jezik. S napretkom tehnologije razvili su se i razni pristupi strojnog prevođenja od kojih je najnoviji neuralno strojno prevođenje (engl. *Neural Machine Translation – NMT*) koji se temelji na dubokom učenju (Lanners, 2019). Implementiran je u Google prevoditelj, jedan od najpoznatijih i najefikasnijih alata za prevođenje.

- Dijaloški sustavi (engl. *Dialogue Systems*) – programi koji komuniciraju s korisnicima prirodnim jezikom, bilo u tekstualnom i/ili govornom obliku. Mogu se podijeliti na sustave koji su orijentirani na izvršavanje zadataka kao što su Siri, Alexa i Cortana te na chatbotove koji su dizajnirani za duže razgovore (Jurafsky & Martin, 2009).

2.2. Područja

Obrada prirodnog jezika se može podijeliti na dva područja: razumijevanje prirodnog jezika (eng. *Natural Language Understanding - NLU*) i generiranje prirodnog jezika (eng. *Natural Language Generation - NLG*) (Kumar, 2018). Sve veća uporaba chatbotova i tehnologije koja uključuje jezik i govor te evolucija poruka od ručnog do automatiziranog unosa pridonosi razvitku sustava za razumijevanje i generiranje prirodnog jezika. Ova područja se međusobno dopunjuju na načina da sustavi za razumijevanje prirodnog jezika određuju značenje podataka na temelju gramatike i konteksta, tekst se zatim pretvara u strukturirane podatke te sustav za generiranje prirodnog jezika uzvraća tekst baziran na danim strukturiranim podacima (Kumar, 2018).

Razumijevanje prirodnog jezika se odnosi na razdvajanje teksta ili govora na manje jedinice koje su lakše razumljive te korištenje sintaktičke i semantičke analize teksta i govora da bi se odredilo značenje rečenice (Dialani, 2020). Fokus je na omogućavanju stroju da razumije ljudski jezik na način da se nestrukturirani podaci organiziraju kako bi ih stroj mogao analizirati i razumjeti. Neki od primjera korištenja ovog tipa obrade prirodnog jezika su: filtriranje neželjenog sadržaja, detekcija sentimenta, određivanje teme sadržaja, strojno prevođenje,

odgovaranje na pitanja, kategorizacija teksta, detekcija entiteta itd. (Kumar, 2018). U prirodnom jeziku često može doći do slučajeva u kojima sličan sadržaj ima različite implikacije, da različite riječi imaju isto značenje ili da se značenje mijenja s određenim kontekstom. Poznavanje standarda i strukture jezika te razumijevanje sadržaja bez višeznačnosti neke su od poteškoća s kojima se suočavaju sustavi za razumijevanje prirodnog jezika.

Generiranje prirodnog jezika se odnosi na proces stvaranja smislenih rečenica na prirodnom jeziku (Kaur, 2021). Pretvara strukturirane podatke dobivene razumijevanjem prirodnog jezika u rečenice koje su lako razumljive za ljude. U početku su sustavi za generiranje prirodnog jezika koristili predloške za generiranje teksta, no oni nadilaze sustave temeljene na predlošcima i generiraju tekst na temelju danih ulaznih podataka. Neki od modela koji su ovo učinili mogućim su (Kaur, 2021):

Markovljev lanac (engl. *Markov chain)*

Markovljev lanac je stohastički proces koji podrazumijeva prijelaze iz jednog stanja u drugo na temelju određenih pravila vjerojatnosti (Soni, 2018). Karakterizira ga "svojstvo zaboravljivosti" (eng. *memoryless property*) što znači da vjerojatnost budućih stanja ovisi samo o sadašnjem stanju, zanemarujući ona koja su mu prethodila (Rocca, 2019). Dakle, da bismo predvidjeli buduće ponašanje procesa, dovoljno je poznavati trenutno ponašanje. Prema tom svojstvu, Markovljev lanac se može primijeniti u diskretnim (prebrojivim ili konačnim) procesima (Rocca, 2019). Primjer toga bio bi bacanje novčića pošto taj proces uključuje diskretne slučajne varijable koje su u različitom trenutku neovisne jedna od drugoj (Rocca, 2019). Ovaj proces se u kontekstu obrade prirodnog jezika može primijeniti u označavanju vrsta riječi (engl. *Part-of-Speech Tagging*) (Tyagi, 2021). To podrazumijeva označavanje svake riječi u rečenici s njezinom odgovarajućom oznakom za vrstu riječi. U Markovljevom lancu svaka riječ predstavlja jedno diskretno stanje te se u tom slučaju može primijeniti na način da se izračuna vjerojatnost pojavljivanja određenih vrsta riječi unutar rečenice koje se nalaze jedna uz drugu. Na primjer, velika je vjerojatnost da će nakon imenice koja se nalazi na početku rečenice uslijediti glagol. Označavanje vrsta riječi omogućuje nam prikupljanje velike količine informacija o riječima što je nužno za ispunjavanje naprednih zadataka obrade prirodnog jezika poput raščlambe, semantičke analize, prijevoda itd.

Ponavljajuća neuronska mreža (engl. *Recurrent neural networks-RNN)*

Ponavljajuća neuronska mreža (RNN) vrsta je napredne umjetne neuronske mreže (ANN) dizajnirane za prepoznavanje sekvencijalnih karakteristika podataka na temelju kojih koristi obrasce za predviđanje scenarija (IBM, 2020). Za razliku od Markovljevog lanca, ovaj model može koristiti ulazne podatke koji su međusobno ovisni te uzima u obzir prethodna stanja za predviđanje. Ponavljajuće neuronske mreže su među najperspektivnijim algoritmima u uporabi jer imaju internu memoriju koja im omogućuje preciznije predviđanje i bolje razumijevanje niza i njegovog konteksta (Donges, N., 2021). Duga kratkoročna memorija (engl. *Long short-term memory,* LSTM) je među najpopularnijim tipovima ponavljajućih neuronskih mreža. Jedinice LSTM-a koriste se kao gradivne jedinice za slojeve RNN-a koji čine LSTM mrežu (Donges, 2021). LSTM-ovi omogućuju RNN-ima da pamte ulazne podatke tijekom dugog vremenskog razdoblja. Ovo je moguće iz razloga što LSTM-ovi sadrže ugrađenu memoriju za pohranu podataka, sličnu kao i memoriji računala te mogu čitati, pisati i brisati podatke iz svoje memorije. Ova mreža pamti informacije duži period bez ponovnog učenja, što čini cijeli taj proces jednostavnijim i bržim. Primjenjuju se u različite svrhe poput: modeliranja jezika, sažimanja teksta, analize sentimenta, strojnog prevođenja, prepoznavanja govora, detekcije lica, prepoznavanja slika, generiranja opisa za slike itd.

Transformeri

Transformeri su neuronske mreže čija se arhitektura zasniva na samopozornosti (engl. *self-attention*) i na principu kodiranja i dekodiranja (engl. *encoder – decoder network*) (Nikulski, 2021). Predstavljeni su 2017. god. u radu Google tima pod nazivom „Attention Is All You Need“ (Vaswani et al., 2017). Ovim radom ostvaren je veliki pomak u korištenju takozvanih „attention“ mehanizama što je i glavni napredak za modele transformera. Ovi mehanizmi omogućuju direktno modeliranje veza između svih riječi u rečenici, bez obzira na njihov položaj. Pomoću toga uzimaju u obzir udaljeniji kontekst određene riječi. Izračunavanje relevantnog konteksta oko riječi se može provoditi paralelno čime se značajno štedi na resursima. Upravo se po tome razlikuju od RNN-a. Naime, iako su i RNN-i i transformeri dizajnirani za obradu sekvencijalnih ulaznih podataka, transformeri ne zahtijevaju obradu tih podataka u određenom redoslijedu (Uszkoreit, 2017). Trenutno najpoznatiji modeli koji se koriste za rješavanje zadataka OPJ-a se sastoje od velikog broja transformera ili njihovih sličnih varijanti. Jedan od najistaknutijih takvih modela je BERT (*Bidirectional Encoder Representations from Transformers*) o kojem će više riječi biti u kasnijem poglavlju.

3. Reprezentativno učenje

Učinkovitost metoda strojnog učenja uvelike ovisi o izboru prikaza podataka na kojem se primjenjuju. Većina podataka u digitalnom okruženju je još uvijek neobrađena (engl. *raw data*), uključujući slike, videozapise, tekst i sl. Iz tog razloga, velik dio napora u strojnom učenju se ulaže u stvaranje reprezentacija podataka na način da ih stroj može naučiti i pomoću toga izvršiti određeni zadatak (Bengio et al., 2013). Primjer toga bio bi postavljanje upita na YouTube-u na temelju kojeg YouTube vraća videozapise koji su najsličniji riječima u korisničkom upitu (Schiappa, 2021). Stoga bi cilj reprezentativnog učenja (engl. *representation learning*) bio omogućiti sustavu da, prema danim neobrađenim podacima, automatski identificira reprezentacije potrebne za otkrivanje relevantnih značajki i klasifikaciju podataka.

Duboko učenje je tipičan pristup reprezentativnog učenja koji se pokazao velikim uspjehom u kontekstu prepoznavanja govora, računalnog vida i obrade prirodnog jezika (Liu et al., 2020). Metode dubokog učenja su zapravo metode reprezentativnog učenja sa više razina reprezentacije (Bengio, 2012). Reprezentacija podataka određuje koliko se korisnih informacija može izvući iz neobrađenih podataka za daljnju klasifikaciju ili predviđanje. Što je više korisnih informacija pretvoreno iz neobrađenih podataka u reprezentacije značajki tih podataka, učinkovitost klasifikacije ili predviđanja će biti bolja (Liu et al., 2020). Generalno govoreći, dobra reprezentacija je ona koja čini naredne zadatke učenja lakšim, a izbor reprezentacija ovisi o tome kakav zadatak želimo izvršiti. Reprezentacije u neuronskim mrežama mogu biti lokalne (engl. *local representation*) ili distribuirane (engl. *distributed representation*) (Liu et al., 2020). Kod lokalne reprezentacije, jedan je entitet predstavljen kao jedna vrijednost, dok je u slučajevima distribuirane reprezentacije svaki entitet predstavljen uzorkom aktivnosti raspoređenim u više elemenata.

3.1. Lokalna reprezentacija

Kako bi se riječi mogle staviti u algoritam strojnog učenja, one trebaju biti pretvorene u vektorske prikaze. Odnosno, riječi se pretvaraju u određene brojeve. Proces pretvaranja teksta u brojeve naziva se vektorizacija (engl. *vectorization*) (Liu et al., 2020). Vektori zatim u međusobnom odnosu tvore vektorski prostor. Najjednostavnija metoda vektorizacije riječi bila bi korištenjem *one-hot* vektora koji sadrži dimenzije danog seta vokabulara, odnosno skupa jedinstvenih riječi iz korištenog skupa podataka (Liu et al., 2020). Na temelju zadanog seta vokabulara pridružuje se 1 odgovarajućem položaju riječi, a 0 drugim riječima. Na ovaj način

one-hot vektori prikazuju riječi samo na način da ih razlikuju jedne od drugih bez ikakvih dodatnih informacija o njihovom značenju. Ipak, to im omogućuje jedinstvenu identifikaciju riječi što znači da dvije riječi neće imati istu vektorsku reprezentaciju.

N-gram modeli se koriste za predviđanje sljedeće riječi u rečenici ili slijeda riječi određivanjem vjerojatnosti njihova pojavljivanja (Jurafsky & Martin, 2009). Naime, ako želimo predvidjeti sljedeću riječ u nekom slijedu, pogledat ćemo prethodne riječi i na temelju velikog korpusa izračunati koja će riječ ili slijed riječi biti najvjerojatniji da se pojavi na toj poziciji. N-gram predstavlja slijed n broja riječi: 2-gram (bigram) je slijed od dvije riječi kao na primjer „sunčan dan“ ili „velik pas“, 3-gram (trigram) je slijed od tri riječi kao na primjer „lijep sunčan dan“ ili „moj velik pas“, dok slijed od četiri riječi ili više nazivamo 4-gram ili jednostavno n-gram (Jurafsky & Martin, 2009). Neki od primjera korištenja ovog modela u OPJ-u su: automatsko dopunjavanje rečenica, automatska provjera pravopisa i prepoznavanje govora.

Bag-of-words (BOW) modeli promatraju dokument kao skup riječi, pritom zanemarujući gramatiku i poredak riječi u rečenici, no uzimajući u obzir učestalost pojedinih riječi unutar dokumenta (Liu et al., 2020). Prema tome se najveći značaj dodjeljuje onoj riječi koja se najviše puta pojavljuje u dokumentu, dok se najmanji značaj dodjeljuje riječi koja se najrjeđe pojavljuje. Ovi modeli su vrlo jednostavni za shvatiti i implementirati te nude veliku fleksibilnost pošto ih je moguće prilagoditi za specifične tekstualne podatke (Liu et al., 2020). Moguće ih je proširiti tako da se koriste bigrami ili trigrami umjesto individualnih riječi te se tako uvaži „inherentna sekvencijalnost jezika“ (Skansi & Lauc 2018). Iako ne uzima u obzir kontekst i značenje riječi pokazao se vrlo uspješnim u problemima predviđanja poput modeliranja jezika i klasifikacije dokumentacije.

Lokalna reprezentacija može biti vrlo korisna u izvršavanju zadataka poput klasifikacije dokumenata, izvlačenja informacija i filtriranja sadržaja. Ipak, takve reprezentacije ne sadrže nikakve semantičke ni sintaktičke informacije o riječima, stoga ne mogu ni prepoznati određene sličnosti u značenju više različitih riječi ili rečenica. Taj problem se nastoji riješiti korištenjem distribuirane reprezentacije.

3.2. Distribuirana reprezentacija

Ideju distribuirane reprezentacije prvi je predstavio Geoffrey E. Hinton 1984. godine u radu koji je kasnije uključen u sadržaj knjige pod nazivom „Parallel Distributed Processing“ (Liu et al., 2020). U knjizi su predstavljene neuronske mreže koje modeliraju ljudske kognitivne procese i inteligenciju čime je inspirirana ideja za distribuiranu reprezentaciju. Glavni koncept u načinu rada ove vrste reprezentacije je da je svaki entitet reprezentiran uzorkom aktivnosti distribuiranom nad mnogo elemenata te je svaki element uključen u reprezentaciju mnogo različitih entiteta. Na osnovi toga, distribuirana reprezentacija gradi prostor sličnosti u kojem semantički slični uzorci ostaju blizu jedni drugima u nekom smislu "udaljenosti" (Theodoridis, 2020). U usporedbi s lokalnom reprezentacijom, distribuirana reprezentacija može prikazati podatke na puno kompaktniji i efikasniji način.

Pojam vektorski prikazi riječi (engl. *word embedding*) podrazumijeva prikaz riječi u obliku vektora na način da slične riječi imaju slično kodirane vektore (Jurafsky & Martin, 2009). To znači da su vektori tih riječi bliže jedan drugome unutar vektorskog prostora. Svaka riječ se preslikava u jedan vektor, a njihove vrijednosti su naučene tehnikama nadziranog učenja (engl. *supervised learning*) poput neuronskih mreža uvježbanih na zadacima ili tehnikama nenadziranog učenja (engl. *unsupervised learning*) kao što je statistička analiza dokumenata (Jurafsky & Martin, 2009). Pomoću vektorskih prikaza riječi nastoji se prikazati semantičko, kontekstualno i sintaktičko značenje svake riječi iz određenog vokabulara. Bitno je naglasiti da individualne dimenzije vektora ne daju puno informacija, međutim, ukupni uzorci udaljenosti i položaja između različitih vektora su vrlo korisni za strojno učenje (Pandey, 2019). Jedan od vodećih modela distribuiranih reprezentacija koji koristi vektorske prikaze riječi kao naučene parametre je Neural Probabilistic Language Model (NPLM) (Liu et al., 2020). Ovaj model prvo dodijeli vektor svakoj riječi, a zatim koristi neuronsku mrežu za predviđanje sljedeće riječi. Na temelju njegovog uzora, razvili su se napredniji modeli za reprezentaciju među kojima je i BERT.

3.2.1. BERT

BERT je tehnika strojnog učenja bazirana na modelu transformera za OPJ koju je razvio Google 2018. god. Predstavljen je u radu „*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*“ u kojem su objašnjeni novi pristupi i tehnike koje koristi BERT (Devlin et al., 2019). Ključna inovacija BERT-a je u primjeni dvosmjernog uvježbavanja (engl. *bidirectional training*) transformera. Za razliku od modela koji su mu prethodili poput

Word2vec-a ili GloVe-a koji generiraju jedan vektorski prikaz riječi za svaku riječ u vokabularu promatrajući tekst samo u jednom smjeru, BERT ih generira uzimajući u obzir riječi koje se nalaze prije i poslije određene riječi, dakle uzima u obzir cijeli kontekst rečenice. BERT je vježban na neoznačenim podatcima preuzetim iz BooksCorpus-a s 800 milijuna riječi i engleske wikipedije s 2 500 milijuna riječi (Devlin et al., 2019).

Dvosmjerno vježbanje transformera bilo je moguće ostvariti korištenjem tehnike modeliranja maskiranog jezika (engl. *masked-language modeling* – MLM) (Devlin & Chang, 2018). Ova tehnika "maskira" neke od ulaznih riječi te zatim koristi ostale riječi u rečenici da bi predvidjela "maskirane" riječi. Iako je MLM prisutan već duže vrijeme, prvi put je pomoću njega uspješno uvježbana duboka neuronska mreža, odnosno BERT (Devlin & Chang, 2018). Ova tehnika je vrlo korisna za vježbanje jezičnih modela bez ljudskog označivanja i posebno korisna za razumijevanje potrebe u korisničkim upitima za pretraživanje. Rezultati Google-ova rada o BERT-u dokazuju da jezični model koji je dvosmjerno vježban ima bolje razumijevanje jezičnog konteksta i tijeka rečenica od jednosmjernih jezičnih modela što je i razlog njegove implementacije u Google-ov sustav za pretraživanje (Devlin et al., 2019). U prosincu, 2019. objavljeno je da je BERT primijenjen na više od 70 jezika u Google pretražitelju, a u listopadu 2020. BERT je obrađivao gotovo svaki pojedinačni upit na engleskom jeziku (Nayak, 2019). Na temelju brojnih testiranja, dokazano je da BERT uspijeva shvatiti suptilne nijanse jezika dajući veću pozornost kontekstualnom okruženju upita (Devlin et al., 2019). Osim MLM-a, BERT je također vježban tehnikom predviđanja sljedeće rečenice (engl. *next sentence prediction* – NSP). Za mnoge zadatke u OPJ-u je potrebno razumijevanje odnosa između različitih rečenica. Tijekom prethodnog vježbanja, ova tehnika spaja dvije maskirane rečenice kao ulazne podatke. Ponekad su to rečenice koje se u izvornom tekstu nalaze jedna uz drugu, a ponekad ne. Ova tehnika tada mora predvidjeti slijede li te dvije rečenice jedna drugu ili ne. Na taj način model uči unutarnju reprezentaciju jezika u procesu prethodnog uvježbavanja koja se kasnije može koristiti za izdvajanje značajki korisnih za izvršavanje zadataka.

4. Projektni rad

Projektni dio ovog rada obuhvaća vizualizaciju sentimentom označenog podatkovnog skupa IMDB recenzija od 4086 rečenica na hrvatskom i engleskom jeziku u Tensorflow's Tensorboard Embedding Projector-u korištenjem pet različitih preduvježbanih modela zasnovanih na BERT-ovoj arhitekturi. Tehnologije korištene za izradu ovog projekta su Google Colaboratory, Hugging Face i Tensorflow's Tensorboard Embedding Projector. Korišten je programski jezik Python iz razloga što ga podržava Google Colaboratory te zato što je to vodeći programski jezik za OPJ zbog svoje jednostavne sintakse, strukture i alata za obradu teksta.

4.1. Skup podataka

U svrhu vizualizacije podataka korištene su recenzije korisnika preuzete s javno dostupnog izvora IMDB-a, najveće baze podataka za informacije o filmovima. Uzet je upravo ovaj skup podataka iz razloga što IMDB sadrži veliku količinu recenzija pisanih prirodnim jezikom i jer su one vrlo pogodne za označavanje sentimenta pošto već sadrže ocjenu koju je korisnik dao. Također, recenzije obično sadrže čvrsto pozitivno ili negativno mišljenje što olakšava označavanje, ali i omogućuje modelima transformera lakše prepoznavanje sentimenta. Važno je napomenuti da rečenice ovog skupa podataka sadrže gramatičke pogreške, kolokvijalne izraze, izostavljene riječi, sarkazam, metafore i slične neformalnosti kojima se tehnologije OPJ-a još uvijek nisu u potpunosti prilagodile. Pošto je jedan od ciljeva bio vizualizirati rečenice na hrvatskom jeziku bilo je potrebno prevesti recenzije s engleskog jezika.

Kao pomoć pri prevođenju korišten je Google prevoditelj. Iako je strojno prevođenje uvelike napredovalo, na temelju korištenja Google prevoditelja zamijećene su različite pogreške u njegovoj izvedbi. Među najčešćima se pokazalo doslovno prevođenje naziva filmova i određenih izraza prisutnih u engleskom jeziku te poteškoće pri korištenju padeža u hrvatskom jeziku:

Rečenica na engleskom jeziku	**Prijevod Google Prevoditelja na hrvatski**
I caught this film late at night on HBO.	Ovaj sam film ulovio kasno navečer na HBO-u.
We came in few minutes late and only saw the end of the opening scene which turned out to be a good thing since it was too intense for a 3 and a 4 year old.	Došli smo za nekoliko minuta kasno i vidjeli smo samo kraj uvodne scene što se pokazalo dobrom stvar jer je bila preintenzivna za 3 i 4 godine.
Too bad it was released nationwide in theaters the same year as "Fear and Loathing" and "Half-Baked."	Šteta što je objavljen širom zemlje u kinima iste godine kao "Strah i gnušanje" i "Napola pečeni".

Tablica 1 - Prijevod rečenica Google prevoditelja

U tablici 1 predstavljeni su primjeri rečenica koji demonstriraju neke od najčešćih uočenih pogrešaka Google prevoditelja. U prvoj rečenici, izraz „I caught this film“ prevedena je doslovno kao „Ovaj sam film ulovio“, dok bi izvorni govornik to preveo kao „Pogledao sam ovaj film“. U drugoj rečenici možemo uočiti nekoliko pogrešaka: doslovan prijevod početka rečenice „We came in few minutes late“ bi se inače prevodilo kao „zakasnili smo nekoliko minuta“ umjesto „Došli smo nekoliko minuta kasno“; pogrešna uporaba padežnih nastavaka u „što se pokazalo dobrom stvar“; neprepoznavanje imenica „a 3 and a 4 year old“. Treći primjer pokazuje nemogućnost prepoznavanja naziva filmova „Strah i prezir u Las Vegasu“ i „Neodoljive budale“.

Iako je rečenice bilo nužno ispravljati, Google prevoditelj se pokazao efikasnim alatom za pomoć pri prevođenju. U većini jednostavnih rečenica producirao je dobre prijevode s visokom gramatičkom točnošću. Uočene pogreške su nijanse u jeziku koje strojno prevođenje još uvijek nije savladalo. Ipak, Google prevoditelj je u većini slučajeva producirao kvalitetne prijevode koji zahtijevaju vrlo malo ili nimalo prepravljanja. Najveća prednost bi zasigurno bila brzina prevođenja. Samostalno prevođenje rečenica bez uporabe Google prevoditelja bi zahtijevalo puno više uloženog vremena.

Nakon prevođenja, skup podatka sadržavao je 4086 rečenica od kojih su 2043 rečenice na engleskom jeziku i 2043 rečenice na hrvatskom jeziku. Svakoj od tih rečenica bilo je potrebno pridružiti odgovarajući sentiment – pozitivan, negativan ili neutralan. Pozitivnih rečenica bilo je 921, negativnih 711, a neutralnih 411.

4.2. Korištena tehnologija

4.2.1. Hugging Face

Hugging Face je poslužitelj otvorenog tipa koji nudi tehnologije OPJ-a, uključujući biblioteku s prethodno uvježbanim transformerima (HuggingFace.co). Biblioteka s transformerima sadrži 10 613 prethodno uvježbanih modela koji izvršavaju zadatke OPJ-a poput klasifikacije, izvlačenja informacija, odgovaranja na pitanja, sažimanja teksta, prijevoda, generiranja teksta itd. Obuhvaćaju 156 prirodnih jezika i 62 arhitekture modela s objedinjenim aplikacijsko-programskim sučeljem. S Hugging Face-a su preuzeti modeli transformera. Hugging Face također podržava efikasnu interoperabilnost između 3 radna okvira za strojno učenje: TensorFlow, Jax i PyTorch. Za potrebe ovog rada korišten je TensorFlow.

4.2.2. TensorFlow

TensorFlow je jedan od najpoznatijih sveobuhvatnih radnih okvira otvorenog koda za strojno učenje (TensorFlow.org). Sadrži fleksibilan sustav alata, biblioteka i resursa zajednice koji omogućava istraživačima i programerima laku izgradnju i implementaciju aplikacija strojnog učenja. TensorFlow podržava uvježbavanje i povezivanje podataka velikih razmjera: učinkovito koristi stotine jakih servera s omogućenim grafičkim procesorima i pokreće prethodno uvježbane modele za korištenje na raznim platformama. Dovoljno je fleksibilan da podrži eksperimentiranje i istraživanje novih modela strojnog učenja te optimizacije na razini sustava. Izvorno su ga razvili istraživači i inženjeri Google Brain tima za potrebe strojnog učenja i istraživanja o dubokim neuronskim mrežama, ali sustav se zbog svoje fleksibilnosti može primijeniti u širokom spektru drugih domena. Radni okvir TensorFlow je moguće pokretati na CPU i na GPU te pruža zbirku alata za razvoj i uvježbavanje modela pomoću Pythona ili JavaScript-a. Osim toga, TensorFlow nudi mogućnost interaktivne vizualizacije vektorskih prikaza riječi pomoću alata TensorBoard. On nudi mogućnost vizualizacije različitih eksperimenata strojnog učenja, podataka visokih dimenzija, grafikona i histograma modela te pomaže u praćenju mjernih podataka, prikazivanju slika, teksta i audio podataka u različitim dimenzijama (Naushad, 2021). Sve od navedenog pomaže u stvaranju novih spoznaja o tome kako model shvaća dane podatke.

Za vizualizaciju skupa podataka ovog rada korišten je *TensorBoard embedding projector*. Moguće ga je koristiti lokalno, instalacijom TensorFlow-a na osobno računalo, a postoji i samostalna verzija na pregledniku gdje korisnici mogu vizualizirati svoje podatke visokih

dimenzija bez potrebe za instalacijom i pokretanjem TensorFlow-a. U ovom projektu korištena je verzija dostupna na pregledniku.

TensorBoard embedding projector se sastoji od 3 izbornika:

- Podatkovni izbornik – koji se koristi za učitavanje ciljanih podataka za promatranje te pokretanje i bojenje podatkovnih točaka
- Inspektorski izbornik – koji se koristi za traženje određenih točaka i gledanje najbližih susjednih točaka
- Izbornik za vizualizaciju – prostor na kojem je prikazana projekcija

Na podatkovnoj traci potrebno je učitati prethodno transformirane rečenice u njihove vektorske prikaze s više dimenzija. Format podatkovnog skupa vektorskih prikaza mora biti u .tsv formatu čime će vrijednosti biti razdvojene tabulatorom. Nakon toga potrebno je učitati metapodatke za te vrijednosti, također u .tsv formatu. Redoslijed vektorskih prikaza riječi i metapodataka bi trebao biti jednak da bi se na taj način mapirale oznake za vizualizaciju.

4.2.3. Google Colaboratory

Google Colaboratory, koji se najčešće naziva „Google Colab“ ili jednostavno „Colab“ istraživački je projekt za izradu prototipa modela strojnog učenja (Google Colaboratory). To je okruženje temeljeno na Jupyter Notebooks softveru, ali za koji nije potrebna instalacija već djeluje na cloud servisima za koje pruža *runtime* potpuno konfiguriran za duboko učenje i besplatan pristup procesorima GPU, CPU i TPU. Podržava programski jezik Python verziju 3.6.9. S obzirom da uvježbavanje modela i učitavanje skupa podataka za uvježbavanje zahtijeva korištenje velike količine računalnih resursa, Google Colaboratory omogućava da taj proces bude brži i kvalitetniji. Google Colaboratory pruža korisnu nadopunu nedostataka računalne memorije osobnih računala koja nisu u stanju procesuirati toliku količinu podataka. Postoji ograničenje u trajanju sesije i veličini korištenih podataka, međutim svaka sesija se može spremiti, a podaci se mogu držati i na eksternim izvorima poput Google Drive-a. Google Colaboratory je u sklopu ovog projekta korišten za izvršavanje koda za transformaciju rečenica pomoću različitih modela transformera.

4.3. Proces vizualizacije podataka

Korišteno je 5 modela transformera zasnovanih na BERT-ovoj arhitekturi. Promatrani su s obzirom na njihovu raspodjelu rečenica u vektorskom prostoru Tensorflow embedding projector-a s posebnim naglaskom na prepoznavanje jezika i sentimenta. Od 5 modela svi su prethodno uvježbani na više jezika, a njih 2 su prilagođeni analizi sentimenta. Cilj je uočiti sličnosti i razlike različitih modela u prepoznavanju značenja rečenica na dva različita jezika – engleskom i hrvatskom jeziku te određivanju sentimenta kako bi se demonstrirala učinkovitost vizualizacije podataka za svaki pojedini model s obzirom na korišten skup podataka. Također su istaknute neke od poteškoća s kojima su se modeli susreli pri vizualizaciji skupa podataka. Korišteni modeli su:

1. bert-base-multilingual-uncased
2. nlptown/bert-base-multilingual-uncased-sentiment
3. EMBEDDIA/crosloengual-bert
4. EMBEDDIA/crosloengual-bert-sentiment
5. classla/bcms-bertic

Nakon prevođenja rečenica i označavanja sentimenta, bilo ih je potrebno pretvoriti u vektorske prikaze riječi da bi ih bilo moguće vizualizirati u vektorskom prostoru. Pretvorba je izvršena unutar Google Colaboratoryja. Prvi korak bio je instalacija modela transformera s *Hugging Face*-a. Nakon toga učitana je datoteka u .tsv formatu koja sadržava rečenice na engleskom i hrvatskom jeziku s označenim odgovarajućim sentimentom. Bilo je potrebno napraviti liste rečenica i označenog sentimenta pošto ih modeli transformera očekuju kao ulazne podatke za generiranje vektorskih prikaza riječi. Prva lista Ana_sentences sadržava rečenice na engleskom jeziku i prevedene rečenice na hrvatskom jeziku, a druga lista Ana_sentiment sadržava označeni sentiment. Skup označenog sentimenta stavljen je u listu dva puta pošto se jedan skup odnosi na rečenice na engleskom jeziku, a drugi na rečenice na hrvatskom jeziku.

```python
import os
import csv
import pandas as pd
df = pd.read_csv("/content/Reviews_parallel.tsv", sep="\t", )
#df.columns = ["text"]


Ana_sentences =[]
Ana_sentiment =[]

for sentence in df['text'].tolist():
  #print(sentence)
  Ana_sentences.append(sentence)

for sentence in df['translated text'].tolist():
  Ana_sentences.append(sentence)

print(Ana_sentences)

for sentence in df['sentiment'].tolist():
  #print(sentence)
  Ana_sentiment.append(sentence) # sentiment for english

for sentence in df['sentiment'].tolist():
  Ana_sentiment.append(sentence) # sentiment for croatian

print(Ana_sentiment)

```

```
['Read the book, forget the movie!', 'I hope this group of film-makers never re-unites.', 'Primary
['negative', 'negative', 'negative', 'positive', 'negative', 'negative', 'positive', 'positive', '
```

Slika 1 – Generiranje skupa rečenica i sentimenata

Nakon što je učitana datoteka s rečenicama i označenim sentimentom te su napravljene liste Ana_sentences i Ana_sentiment, moguće je pokrenuti model transformera da listu Ana_sentences pretvori u listu vektorskih prikaza riječi. U navedenom kodu korišten je model bert-base-multilingual-uncased:

```python
def run_transformer(model_name = "bert-base-multilingual-uncased"):

    # Load AutoModel from huggingface model repository

    # model_name = "bert-base-multilingual-uncased"
    tokenizer = AutoTokenizer.from_pretrained(model_name)
    model_config = AutoConfig.from_pretrained(
        model_name, output_hidden_states=True)
    model = AutoModel.from_pretrained(model_name, config=model_config)
    model.to("cuda")
    extracted_features = []
    training_generator = torch.utils.data.DataLoader(Ana_sentences, batch_size=512, num_workers=4)
    for batch in training_generator:
      print(batch)

      # Tokenize sentences
      encoded_input = tokenizer(batch, padding="max_length", truncation=True, max_length=128, return_tensors='pt')
      encoded_input.to("cuda")
      with torch.no_grad():
        model_output = model(**encoded_input)
        extracted_features.extend(model_output.pooler_output.detach().cpu().numpy().tolist())
    # Compute token embeddings
    output_folder_name = model_name.split("/")[-1]
    with open(f'/content/All_sentences-{output_folder_name}.tsv', 'w', newline='') as f_output:
      tsv_output = csv.writer(f_output, delimiter='\t')

      for vector in extracted_features:
        tsv_output.writerow(vector)

```

Slika 2 – Inicijalizacija modela i pretvorba podataka u vektorske prikaze riječi

Funkcija run_transformer prima naziv modela kao parametar i potom učitava istoimeni prethodno uvježbani model koji se zatim koristi za pretvaranje skupa podataka u vektorske prikaze riječi.

Nakon dobivenih vektorskih prikaza riječi potrebno je pridružiti im metapodatke kako bi mogli vizualizirali rečenice prema jeziku i sentimentu u radnom okviru Tensorflow's Tensorboard Embedding Projector. Izdvojene su rečenice, sentiment i jezik kao metapodatci kao što je vidljivo na slici 3.

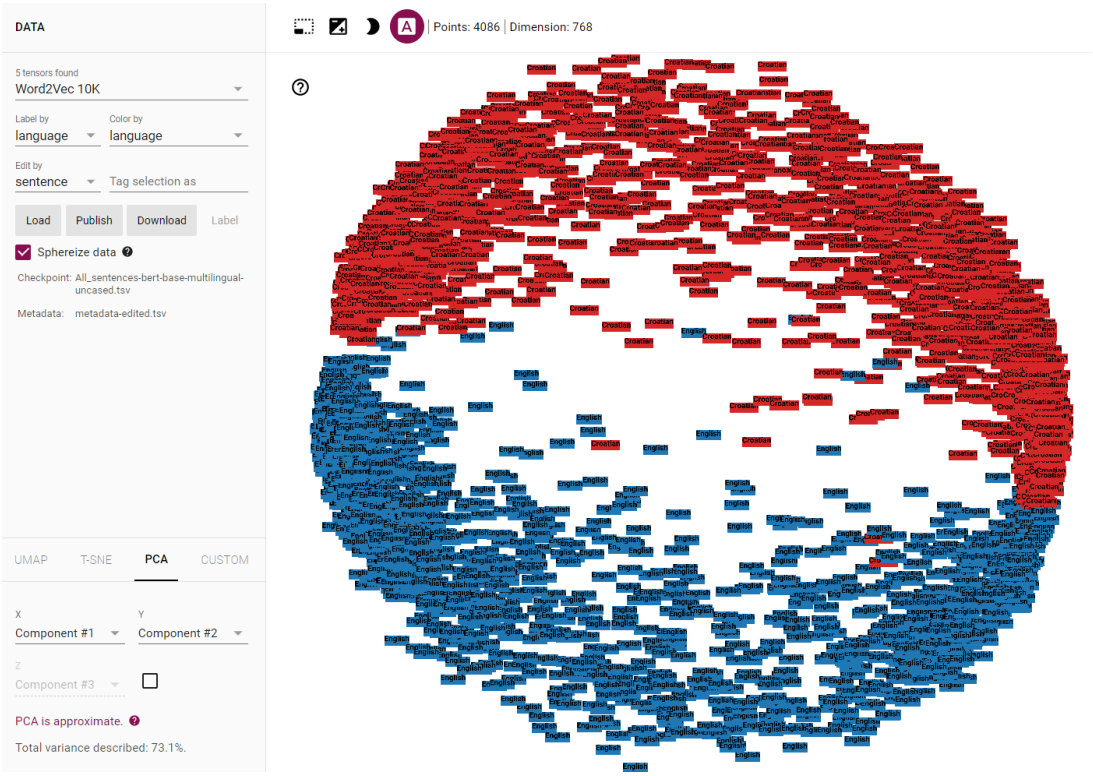
```
1 textfile = open("/content/all_meta.tsv", "w")
2 max = len(Ana_sentences)
3 textfile.write('sentence' + '\t' + 'sentiment' + '\t' + 'language' + '\n')
4 for i, meta in enumerate(Ana_sentences):
5   language = 'English' if (i < max/2) else 'Croatian'
6   textfile.write(meta + '\t' + Ana_sentiment[i] + '\t' + language + '\n')
7 textfile.close()
```

Slika 3 – Izvoz metapodataka

4.4. Modeli

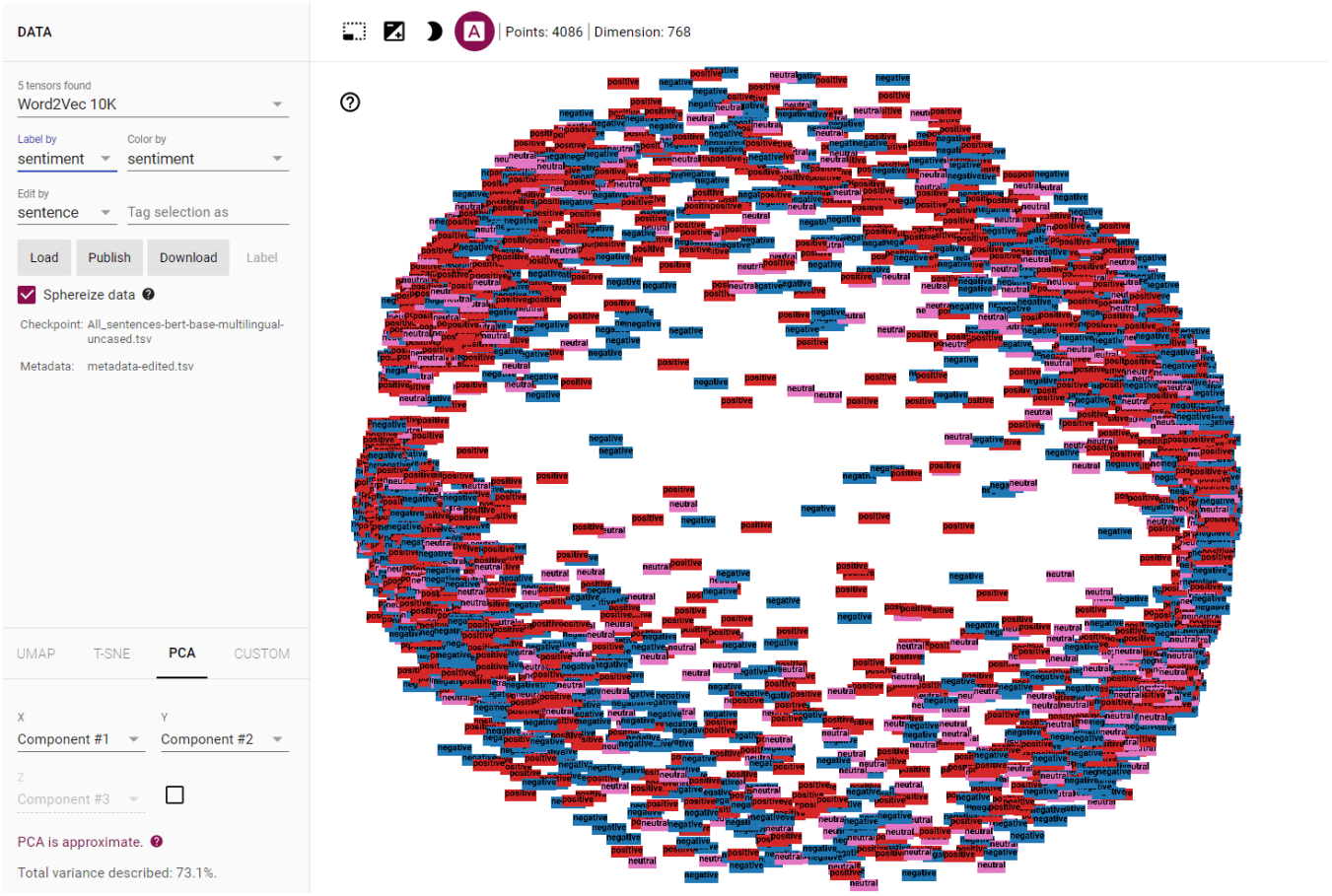
4.4.1. bert-base-multilingual-uncased

Model bert-base-multilingual-uncased je prethodno uvježban na 102 jezika, uključujući hrvatski, koristeći tehniku modeliranja maskiranog jezika (Devlin et al., 2019). Ovaj model je „uncased“ što znači da ne uzima u obzir razlike između velikih i malih slova. Model je prvenstveno usmjeren na prilagodbu (engl. *fine-tuning*) zadataka koji koriste cijelu rečenicu za donošenje odluka poput klasifikacije nizova, klasifikacije tokena ili odgovaranja na pitanja. Korištenjem vektorskih prikaza riječi dobivenih ovim modelom vizualizirana je njihova raspodijeljenost u vektorskom prostoru TensorBoard embedding projector-a:



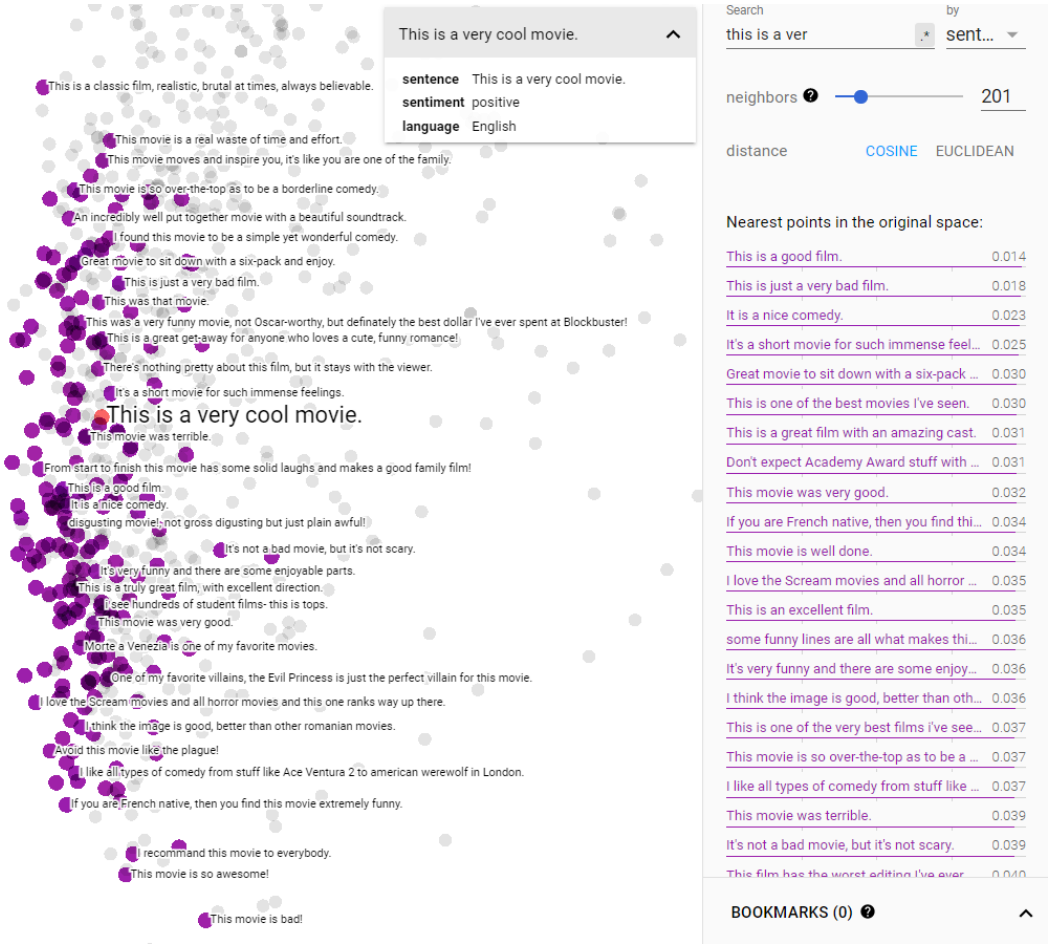
Slika 4 - Vizualizacija skupa podataka po jeziku modelom bert-base-multilingual-uncased

Na slici 4 demonstriran je dvodimenzionalni prikaz raspodjele rečenica označenih i obojanih po jeziku. Crvenom bojom su označene rečenice na hrvatskom jeziku, a plavom na engleskom. Ovaj višejezični model uspijeva vizualno jasno razdvojiti rečenice dva različita jezika s vrlo malim preklapanjem engleskih i hrvatskih rečenica. Ako istom skupu rečenica promijenimo oznake na sentiment, vizualizacija će izgledati ovako:



Slika 5 - Vizualizacija skupa podataka po sentimentu modelom bert-base-multilingual-uncased

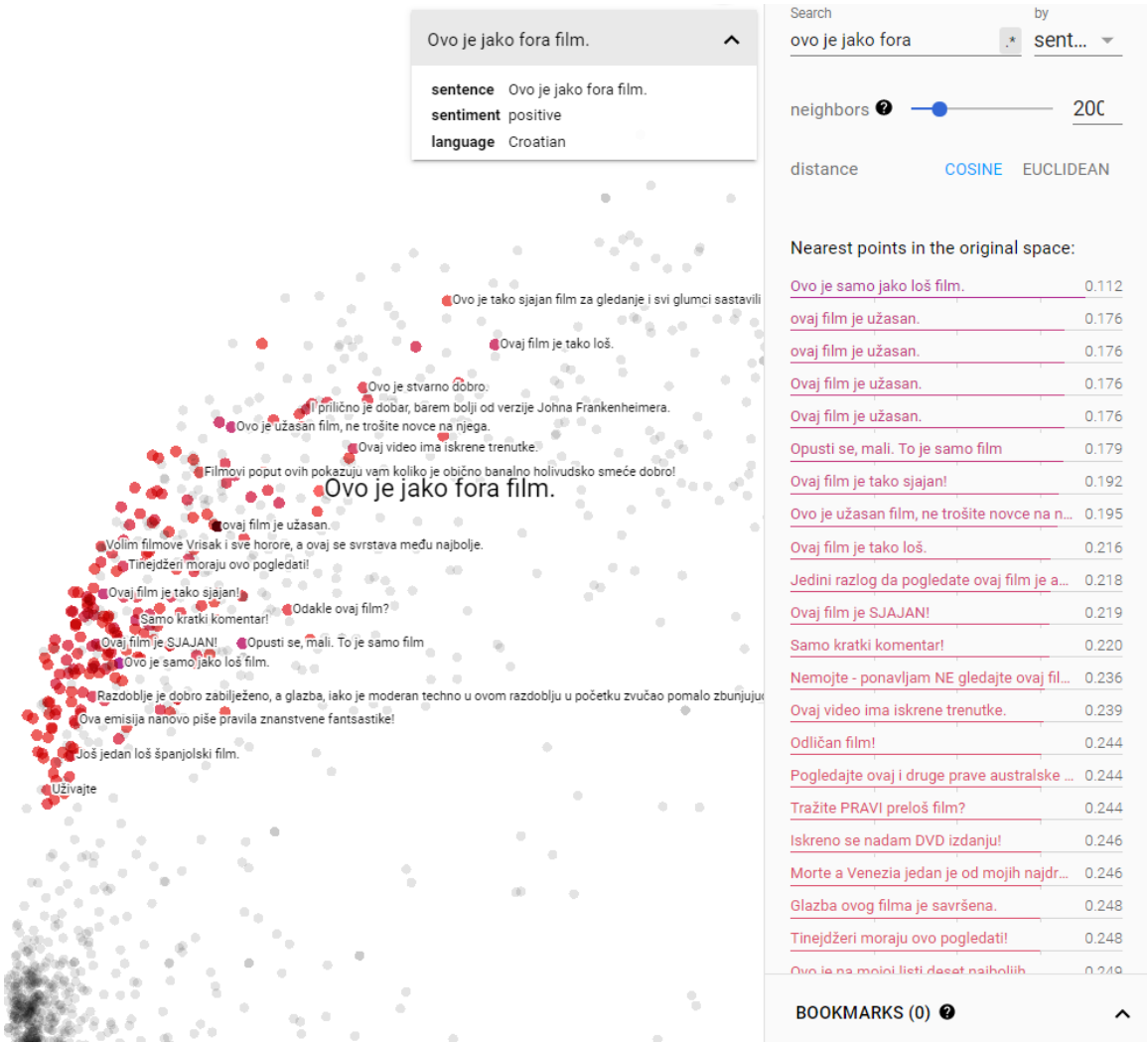
Crvenom su bojom označene pozitivne rečenice, plavom negativne, a ružičastom neutralne. Iako bi model bert-base-multilingual-uncased trebao moći izvući značenje rečenica tehnikom modeliranja maskiranog jezika, nije uspješan u razdvajanju rečenica s obzirom na sentiment pošto nije prilagođen analizi sentimenta. Iz tog razloga ne uspijeva vizualno grupirati rečenice istog sentimenta. Klikom na vektorske prikaze, možemo vidjeti koje rečenice je model prepoznao kao slične njima s obzirom na značenje:



Slika 6 - Primjer rečenice na engleskom jeziku za model bert-base-multilingual-uncased

Klikom na rečenicu „This is a very cool movie.“, pojavljuje se inspektorski izbornik na kojem je vidljivo koje su joj rečenice najbliže u vektorskom prostoru, odnosno koje rečenice je model prepoznao kao najsličnije po značenju označene rečenice. Iako je većina najsličnijih rečenica pozitivnog sentimenta kao i označena rečenica, primjećujemo da su među njima pronađene i one s negativnim sentimentom. Kao npr. rečenice „This is just a very bad film“ i „This movie was terrible“. Ipak, navedene rečenice s negativnim sentimentom sadrže sličnu sintaksu te iste ili slične ključne riječi kao i označena rečenica. Od 100 izoliranih najsličnijih rečenica, njih 60 je imalo pozitivan sentiment, negativan 29, a neutralan 11.

Na isti način možemo pronaći prevedenu rečenicu na hrvatski jezik i vizualizirati njoj slične rečenice:



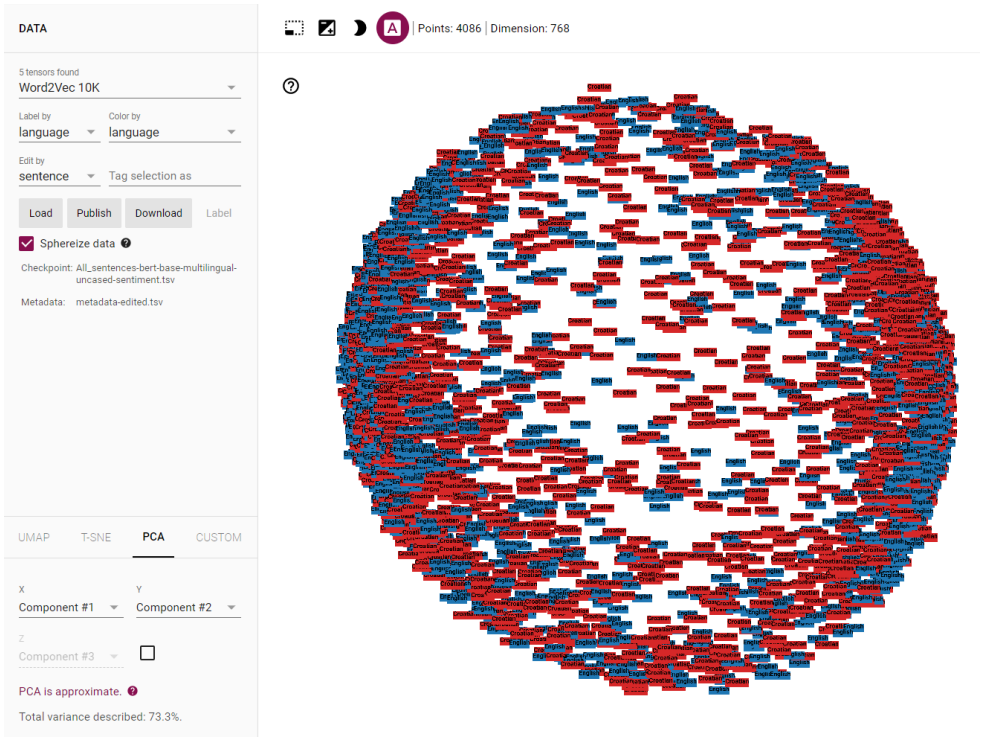
Slika 7 - Primjer rečenice na hrvatskom jeziku za model bert-base-multilingual-uncased

Na slici primjećujemo da su rečenici „Ovo je jako fora film.“ najbliže rečenice u značenju one koje imaju također sličnu strukturu. Međutim, za razliku od rečenice na engleskom jeziku, njoj najbliže rečenice imaju negativan sentiment. Kao npr. rečenice „Ovo je samo jako loš film.“ i „Ovaj film je užasan.“ Od izoliranih 100 najsličnijih rečenica njih 52 su bile pozitivne kao i označena rečenica, 28 negativne i 20 neutralne.

Model bert-base-multilingual-uncased se pokazao uspješnim u prepoznavanju razlike između dva različita jezika. Na temelju primjera iste rečenice na hrvatskom i engleskom jeziku može se zaključiti da model prepoznaje slične strukture rečenica u oba jezika te ključne riječi, što mu je glavni fokus u otkrivanju značenja. Model ne uspijeva vizualizirati razliku u sentimentu rečenica. Iako je u oba primjera pronašao najveći broj sličnih rečenica istog sentimenta, među rečenicama s najbližim značenjem vidljive su rečenice s negativnim sentimentom što je više zastupljeno kod rečenica na hrvatskom nego na engleskom jeziku.

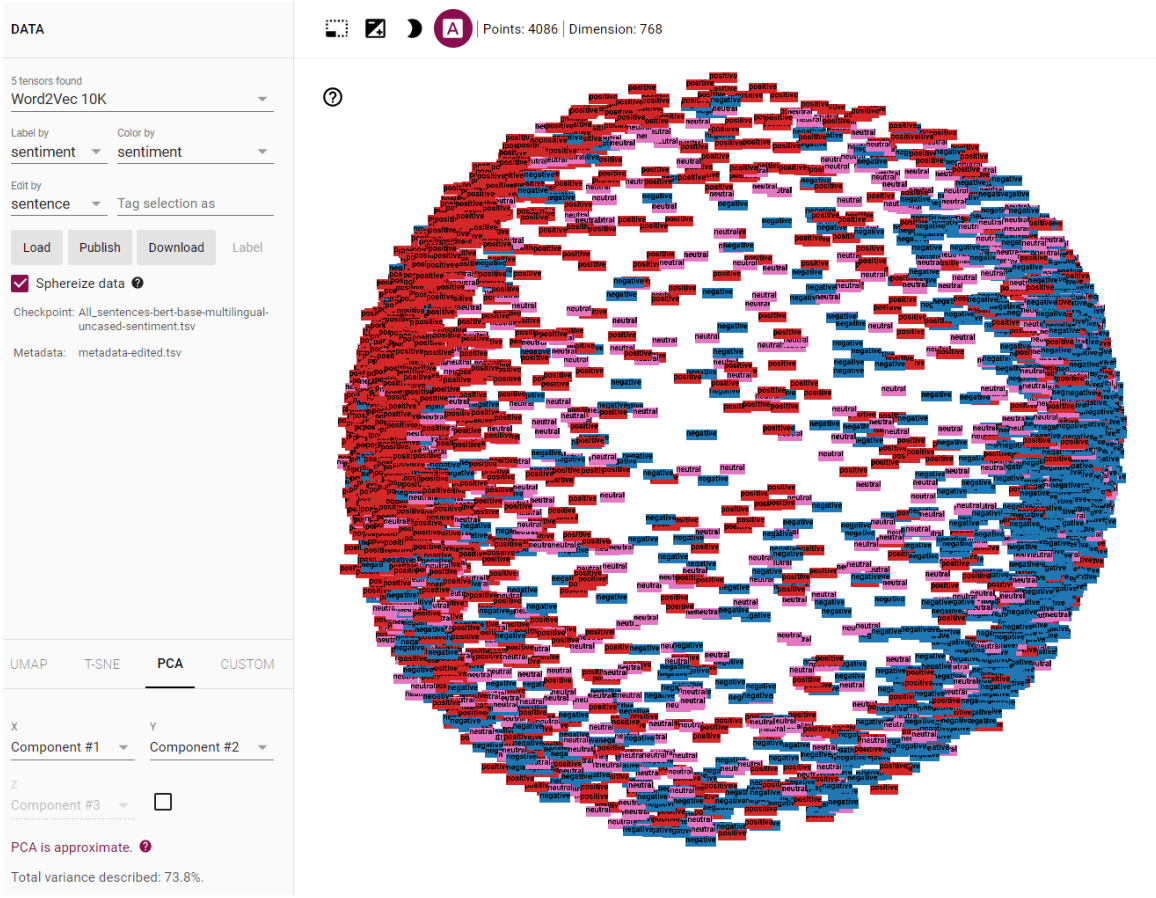
4.4.2. nlptown/bert-base-multilingual-uncased-sentiment

Ovaj model je bert-base-multilingual-uncased model prilagođen analizi sentimenta na 6 jezika: engleskom, nizozemskom, njemačkom, francuskom, španjolskom i talijanskom. Predviđen je za korištenje pri analizi sentimenta za recenzije proizvoda na prethodno navedenim jezicima. Ako promatramo raspodijeljenost rečenica na hrvatskom i engleskom jeziku u vektorskom prostoru označenih i obojanih prema jeziku, vizualizacija izvedbe ovog modela će izgledati kao na slici 8.



Slika 8 - Vizualizacija skupa podataka po jeziku modelom nlptown/bert-base-multilingual-uncased-sentiment

U vizualizaciji rečenica prema jeziku vidimo da postoji malo veća koncentracija crvene boje na desnoj strani, no primjećujemo da model ne vrši jasnu raspodjelu vektorskih prikaza riječi s obzirom na jezik. Pošto je prilagođen analizi sentimenta, model bi trebao grupirati slično značenje rečenica s obzirom na njihov sentiment. Ako promijenimo boju i oznake vektorskih prikaza riječi na sentiment, njihova vizualizacija će izgledati kao na slici 9.



Slika 9 - Vizualizacija skupa podataka po sentimentu modelom nlptown/bert-base-multilingual-uncased-sentiment

Na slici 9 plavom su bojom označene rečenice negativnog sentimenta, crvenom pozitivnog i ružičastom neutralnog. Prema vizualizaciji vektorskih prikaza riječi na slici, može se primijetiti veća koncentracija crvene boje na lijevoj strani te plave na desnoj. Na prikazu rečenica na oba jezika model ih uspijeva vizualno razdvojiti po značenju na temelju sentimenta, no može se uočiti i velik broj pogrešno grupiranih rečenica. Vidljive su pozitivno označene rečenice među negativno označenim rečenicama i obratno. Kako bi mogli vidjeti u kojim slučajevima model griješi, možemo izdvojiti i promotriti rečenice koje se prema sentimentu nalaze u pogrešnim grupacijama.

	Hrvatske rečenice	Engleske rečenice
Pozitivne među negativnima	Pa, nemam puno toga za reći o ovom filmu, osim da je to zaista bio predivan film.	Tragic, while at the same time, absurdly entertaining.
Negativne među pozitivnima	Dobra radnja, dobra režija, loša izvedba!	The special effects are "great" also.

Tablica 2 - Primjeri pogrešno grupiranih rečenica modela nlptown/bert-base-multilingual-uncased-sentiment

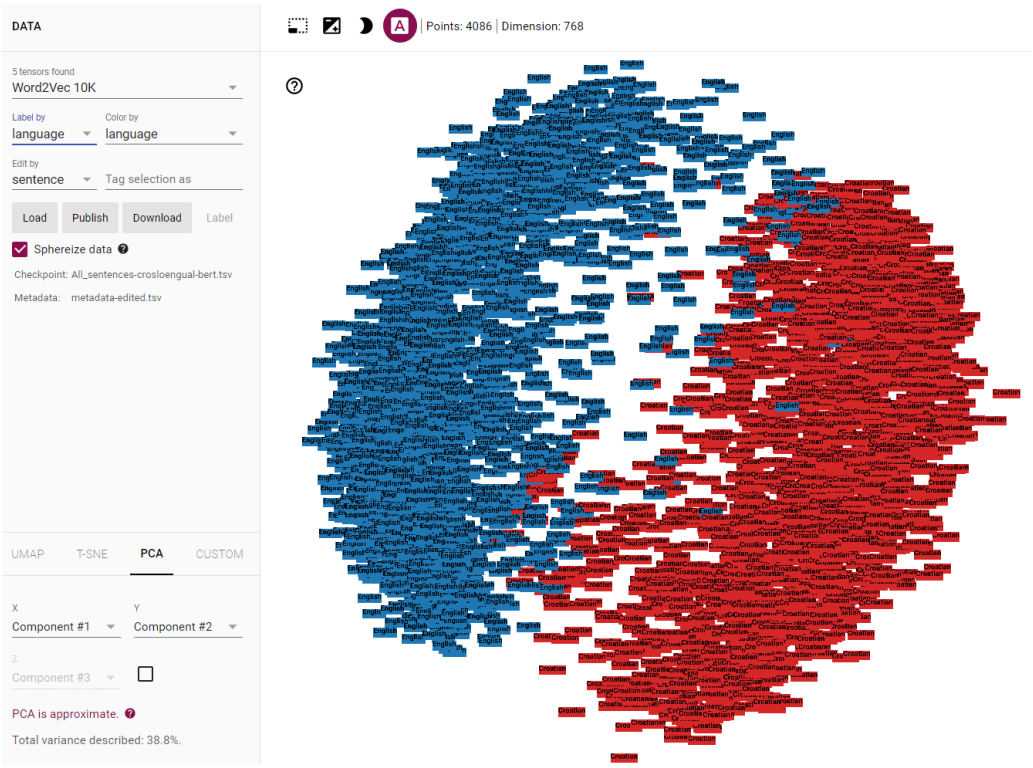
U tablici su dani primjeri rečenica na hrvatskom i engleskom jeziku od kojih je svaka označena pozitivnim ili negativnim sentimentom, ali se prema modelu nalazi najbliže rečenicama drugačijeg sentimenta. Među pozitivno označenim rečenicama na oba jezika, možemo primijetiti da, iako imaju sveukupno pozitivan sentiment, sadrže dijelove koji imaju negativnu ili neutralnu konotaciju. Oni se u obje rečenice nalaze na njihovom početku: „Pa, nemam puno toga za reći o ovom filmu“ i „Tragic“. Ovo je također vidljivo u negativno označenoj hrvatskoj rečenici u kojoj je svukupni dojam negativan, ali sadrži više pozitivnih konotacija: „Dobra radnja, dobra režija“. U zadnjem primjeru negativno označene engleske rečenice, primjećujemo da model ne uspijeva prepoznati navodnike kao sarkazam.

Model nlptown/bert-base-multilingual-uncased-sentiment raspoređuje rečenice po značenju s obzirom na sentiment. Iz tog razloga ne čini jasnu razliku u vizualizaciji između dva različita jezika, ali uspješno razdvaja rečenice s obzirom na sentiment. U istaknutim primjerima rečenica s pozitivno označenim sentimentom kod kojih je model pogriješio, mogu se uočiti dijelovi rečenica s negativnom konotacijom. Isto tako, u primjeru negativno označene hrvatske rečenice, uočavamo dijelove s pozitivnom konotacijom. Vidimo da model u ovim slučajevima ne prepoznaje sveukupan sentiment rečenice, već se fokusira na pojedinačne dijelove koji mu otežavaju određivanje sentimenta. Osim toga, u zadnjem primjeru je vidljivo da model ima poteškoće u prepoznavanju sarkazma u rečenici.

4.4.3. EMBEDDIA/crosloengual-bert

EMBEDDIA/crosloengual-bert je trojezični model, vježban na hrvatskim, slovenskim i engleskim tekstovima (Ulčar & Robnik-Šikonja, 2020). Ovo je jedan od samo dva javno dostupna BERT modela koji je vježban na hrvatskom i jedini koji je vježban na slovenskom i hrvatskom jeziku. Vrlo mala zastupljenost hrvatskog i slovenskog jezika u tehnologijama OPJ-a te manjak javno dostupnih podataka za vježbanje modela na tim jezicima čine ovaj model jako vrijednim resursom za zajednicu istraživača u području OPJ-a.

Na primjeru modela bert-base-multilingual-uncased ustanovljeno je da modeli koji nisu prilagođeni analizi sentimenta, neće moći raspoznati razliku u značenju na temelju toga, no vrlo su uspješni u vizualizaciji na temelju značenja u dva različita jezika. Vizualizacija svih rečenica skupa podataka označenih i obojanih po jeziku modela EMBEDDIA/crosloengual-bert izgledat će kao na slici 10:



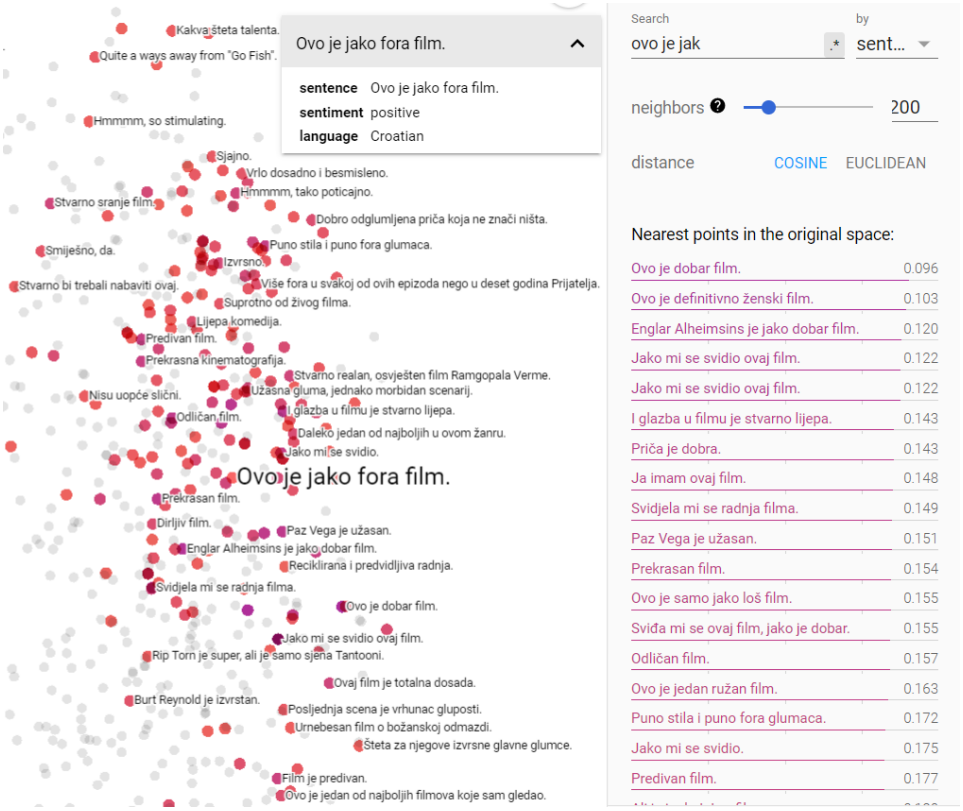
Slika 10 - Vizualizacija skupa podataka po jeziku modelom EMBEDDIA/crosloengual-bert

Na slici se vidi jasna razdvojenost hrvatskog i engleskog jezika s malim preklapanjima. Na temelju istih primjera rečenica korištenih za model bert-base-multilingual-uncased možemo usporediti sličnosti i razlike djelovanja ova dva modela pri određivanju sličnosti u značenju hrvatskih i engleskih rečenica.

Slika 11 - Primjer rečenice na engleskom jeziku za model EMBEDDIA/crosloengual-bert

Na slici 11 prikazane su rečenice koje je model prepoznao kao najsličnije rečenici „This is a very cool movie.“ Među najbližim rečenicama po značenju model pronalazi one sa sličnom sintaksom te sličnim ili istim ključnim riječima. Također, najsličnije rečenice sadrže pozitivan sentiment kao i označena rečenica. Kad izoliramo prvih 100 najbližih rečenica, njih 59 ima pozitivan sentiment, 33 negativan i njih 10 neutralan.

Isto tako, možemo promotriti najbliže rečenice po značenju na temelju iste rečenice prevedene na hrvatski jezik:



Slika 12 - Primjer rečenice na hrvatskom jeziku za model EMBEDDIA/crosloengual-bert

Primjećujemo da model rečenici „Ovo je jako fora film.“ pronalazi po značenju najsličnije rečenice sličnih struktura, ali i sentimenta. Za razliku od modela bert-base-multilingual-uncased koji je kao najsličnije rečenice prepoznao one s negativnim sentimentom, ovaj model uspjeva bolje prepoznati sličnosti u semantičkom značenju navedenog primjera rečenice na hrvatskom jeziku. Od izoliranih 100 najsličnijih rečenica, njih 62 su pozitivne, 29 negativne i 10 neutralne.

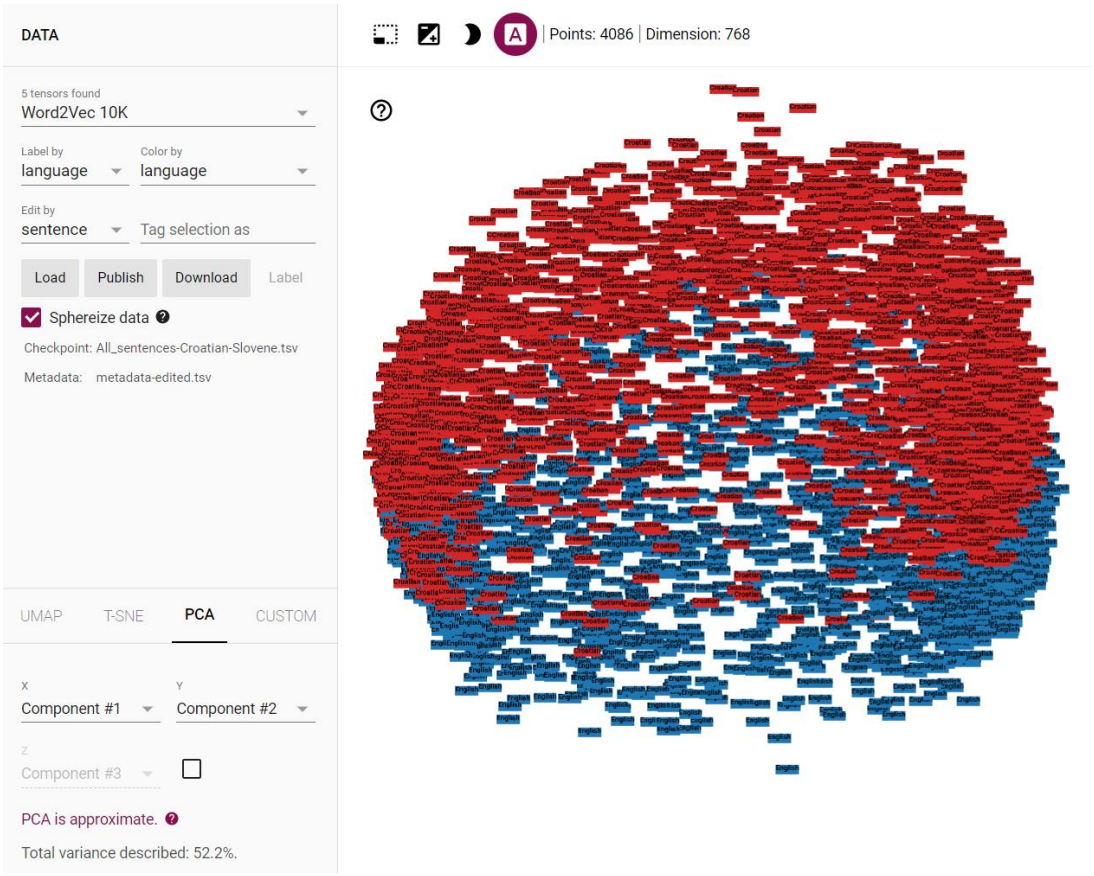
Model EMBEDDIA/crosloengual-bert jasno vizualizira podjelu rečenica na temelju hrvatskog i engleskog jezika. U primjerima rečenica se ne vidi veća razlika u pronalasku broja najsličnijih 100 rečenica s obzirom na sentiment u odnosu na model bert-base-multilingual-uncase.

Međutim, vidljiva je razlika u pronalaženju najsličnijih rečenica na hrvatskom jeziku s obzirom na semantičko značenje. Fokusirajući se na tri jezika, ovaj model ima bolju izvedbu od višejezičnog modela bert-base-multilingual-uncased s obzirom na prepoznavanje značenja u hrvatskim rečenicama.

4.4.4. EMBEDDIA/crosloengual-bert-sentiment

Ovaj model je baziran na EMBEDDIA/crosloengual-bert modelu i vježban na analizi sentimenta (Thakkar, 2021). Korišteni su tekstovi novinskih članaka na slovenskom i hrvatskom jeziku kao skup podataka za vježbanje modela kako bi se poboljšala analiza sentimenta nad hrvatskim tekstovima. Pošto slovenski i hrvatski pripadaju skupini južnoslavenskih jezika, imaju vrlo visoku razinu međusobne razumljivosti. Iz tog razloga je bilo moguće koristiti kontekstualne informacije iz slovenskog skupa podataka za prijenos znanja između dva jezika.

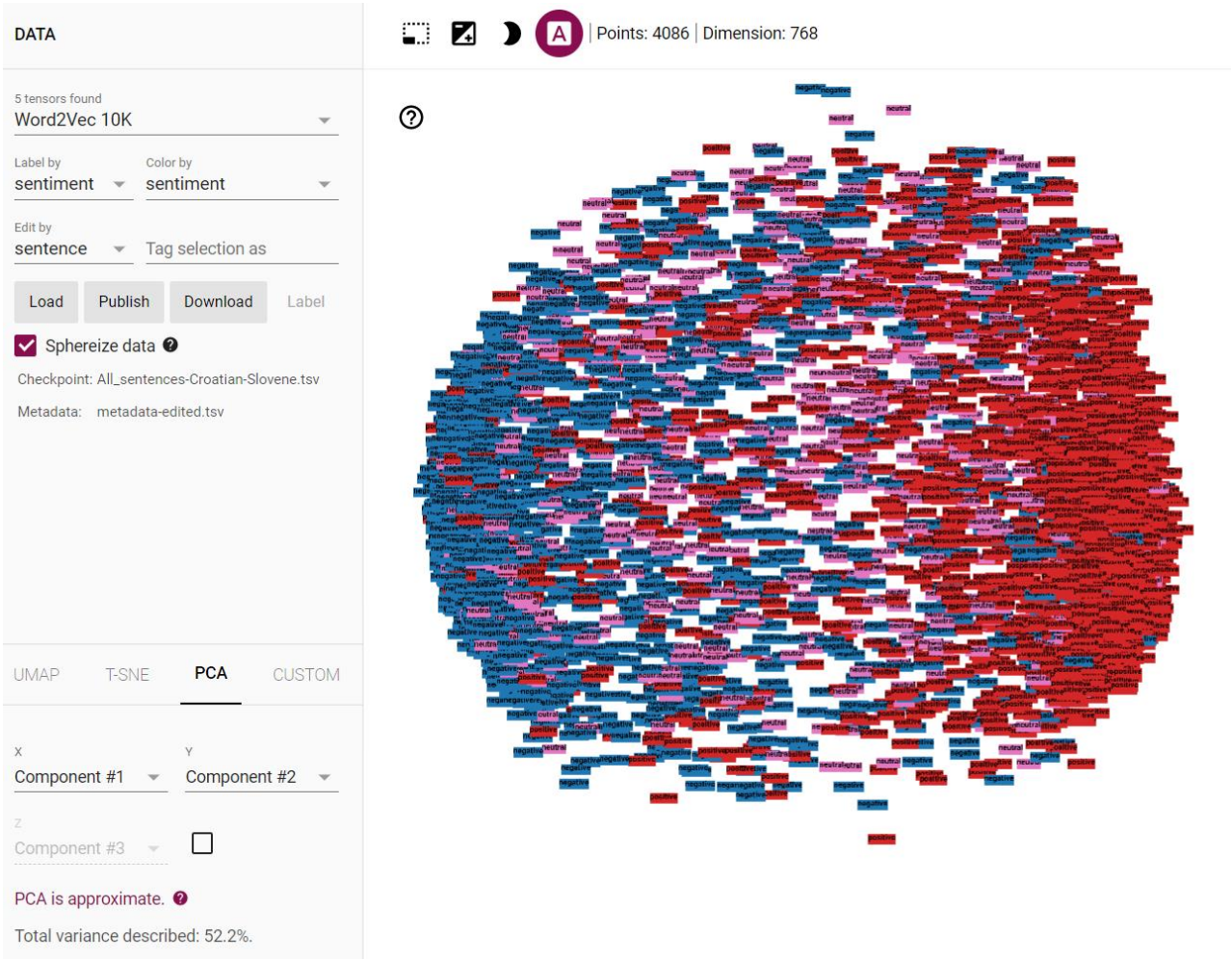
Vizualizacija vektorskih prikaza riječi ovog modela s obzirom na jezik izgleda ovako:



Slika 13 - Vizualizacija skupa podataka po jeziku modelom EMBEDDIA/crosloengual-bert-sentiment

Iako se engleske i hrvatske rečenice na slici 13 međusobno preklapaju, može se uočiti da model stvara razliku između jezika. Usporedno s modelom nlptown/bert-base-multilingual-uncased-sentiment ovaj model bolje uspijeva vizualizirati podjelu rečenica s obzirom na različite jezike. Međutim, treba uzeti u obzir da je ovaj model prilagođen hrvatskom i slovenskom jeziku dok je nlptown/bert-base-multilingual-uncased-sentiment prilagođen na 6 jezika od kojih nijedan nije iz skupine južnoslavenskih jezika.

S obzirom na podjelu prema sentimentu, model će vizualizirati rečenice na sljedeći način:



Slika 14 - Vizualizacija skupa podataka po sentimentu modelom EMBEDDIA/crosloengual-bert-sentiment

Prema raspodijeli rečenica po sentimentu, može se uočiti veća koncentracija crvene boje – pozitivnog sentimenta – na desnoj strani i plave boje – negativnog sentimenta – na lijevoj strani vektorskog prikaza. Može se uočiti vizualna razdvojenost vektorskih prikaza riječi s obzirom na sentiment, međutim, primjećujemo da se neke rečenice nalaze u grupaciji rečenica s drugačijim sentimentom.

U tablici 3 izdvojeni su primjeri rečenica na hrvatskom i engleskom jeziku koje je model grupirao zajedno s rečenicama drugačijeg sentimenta od onoga kojim su označene.

	Hrvatske rečenice	Engleske rečenice
Pozitivne među negativnima	Što god da kažem neće biti dovoljno dobro za ovaj podcijenjen film.	One of the most underrated comedies.
Negativne među pozitivnima	Ovaj je film bio grozan ljubiteljima znanstvene fantastike.	Go get a root canal instead – you'll enjoy it more.

Tablica 3 - Primjeri pogrešno grupiranih rečenica modela EMBEDDIA/crosloengual-bert-sentiment

Od pozitivno označenih rečenica koje je model svrstao među negativno označene, izdvojene su hrvatska i engleska rečenica sa sličnim značenjem. U oba slučaja se pojavljuje riječ „podcijenjen“ ili „underrated“ koja označava da nešto nije dobilo onoliko priznanja koliko zaslužuje. Čini se da model tumači ovu riječ kao riječ obilježenu negativnim sentimentom. Osim toga, u hrvatskoj rečenici se nalazi negacija „neće biti dovoljno dobro“ koja dodatno daje prednost negativnom sentimentu. Primjer negativne hrvatske rečenice koju je model svrstao među pozitivno označene sadrži riječi „grozan“ koja ima negativnu konotaciju i „ljubiteljima“ koja ima pozitivnu konotaciju. U ovom slučaju, model daje prednost pozitivnoj konotaciji. Drugi primjer rečenice na engleskom jeziku ponovno pokazuje nerazumijevanje sarkazma.

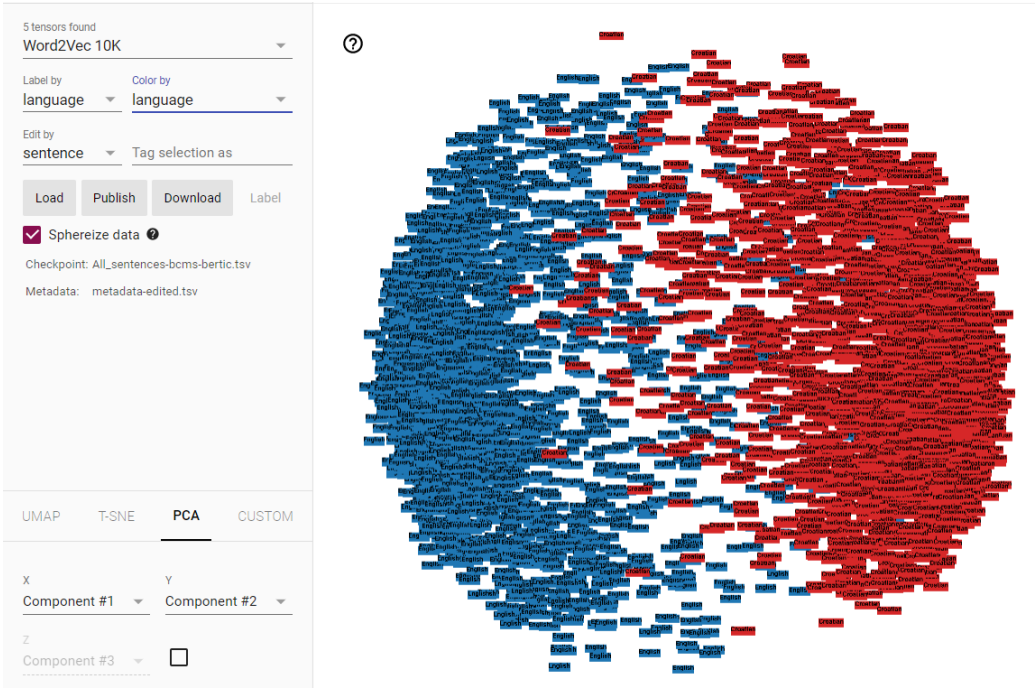
Model EMBEDDIA/crosloengual-bert-sentiment se pokazao uspješnim u vizualizaciji vektorskih prikaza riječi s obzirom na njihovu podijeljenost prema jeziku i sentimentu. Naravno, podjela s obzirom na jezik nije u potpunosti jasna kao što je u slučaju modela bert-base-multilingual-uncased, no svejedno uspijeva vizualizirati razliku između jezika. S obzirom na sentiment, sveukupno je vidljiva odvojenost različitih sentimenta. Ipak, kao i u slučaju modela bert-base-multilingual-uncased-sentiment, ne prepoznaje sarkazam te ima poteškoće u prepoznavanju sentimenta u rečenicama koje sadrže riječi pozitivnih i negativnih konotacija.

4.4.5. classla/bcms-bertic

Model classla/bcms-bertic je prethodno uvježban na tekstovima na bosanskom, hrvatskom, srpskom i crnogorskom jeziku (Ljubešić & Lauc, 2021). Samo ime modela – BERTić – inspirirano je činjenicom da se u hrvatskom jeziku često koristi nastavak -ić za tvorbu umanjenica. Osim toga, u svim zemljama, gdje se govore prethodno navedeni jezici, vrlo su učestala prezimena koja završavaju na sufiks -ić. Za treniranje BERTić-a korišten je ELECTRA pristup koji se pokazao učinkovitijim od BERT modela.

ELECTRA *(Efficiently Learning an Encoder that Classifies Token Replacements Accurately*) je nova metoda prethodnog uvježbavanja modela koja nadmašuje postojeće tehnike na temelju istih računalnih resursa (Clark & Luong, 2020).. Za razliku od BERT-a koji koristi tehniku modeliranja maskiranog jezika za vježbanje modela, ova metoda se temelji na tehnici detekcije zamijenjenih tokena (engl. *replaced token detection* – RTD). Korištenjem MLM-a, u slučaju BERT modela, maskira se 15% tokena, odnosno riječi, na temelju kojih model nastoji predvidjeti koje su to riječi koristeći kontekst u kojem se nalazi. Umjesto da ošteti podatke njihovim maskiranjem, RTD mijenja tokene s pogrešnim, ali mogućim rješenjima. Zadatak prethodnog uvježbavanja zahtijeva da model zatim utvrdi koji su tokeni iz izvornog unosa zamijenjeni, a koji su ostali isti. Ovaj zadatak binarne klasifikacije primjenjuje se na svaki ulazni token, umjesto na samo mali broj maskiranih tokena kao što to radi BERT. To čini RTD učinkovitijim od MLM-a pošto ELECTRA može vidjeti manje primjera za postizanje iste izvedbe. Istodobno, RTD rezultira snažnim reprezentativnim učenjem jer model mora naučiti točan prikaz distribucije podataka kako bi riješio zadatak.

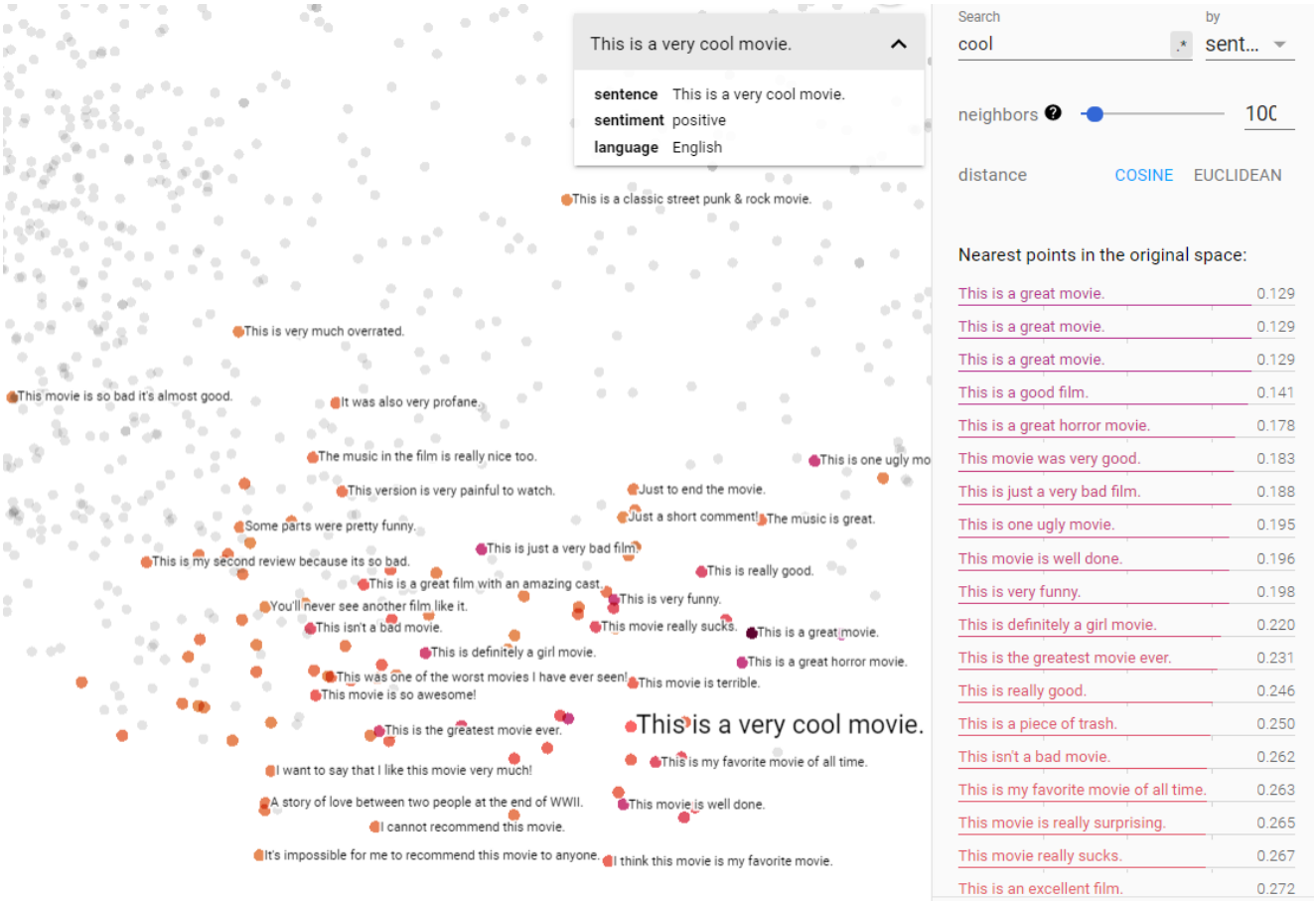
Vizualizacija vektorskih prikaza riječi modela BERTić s obzirom na jezik izgledat će ovako:



Slika 15 - Vizualizacija skupa podataka po jeziku modelom classla/bcms-bertic

Model classla/bcms-bertic uspješno vizualizira rečenice s obzirom na različitost u jeziku s malim preklapanjima. Nije prethodno uvježbavan na analizi sentimenta, stoga nije u mogućnosti vizualizirati razliku u sentimentu rečenica kao što je slučaj s modelima bert-base-

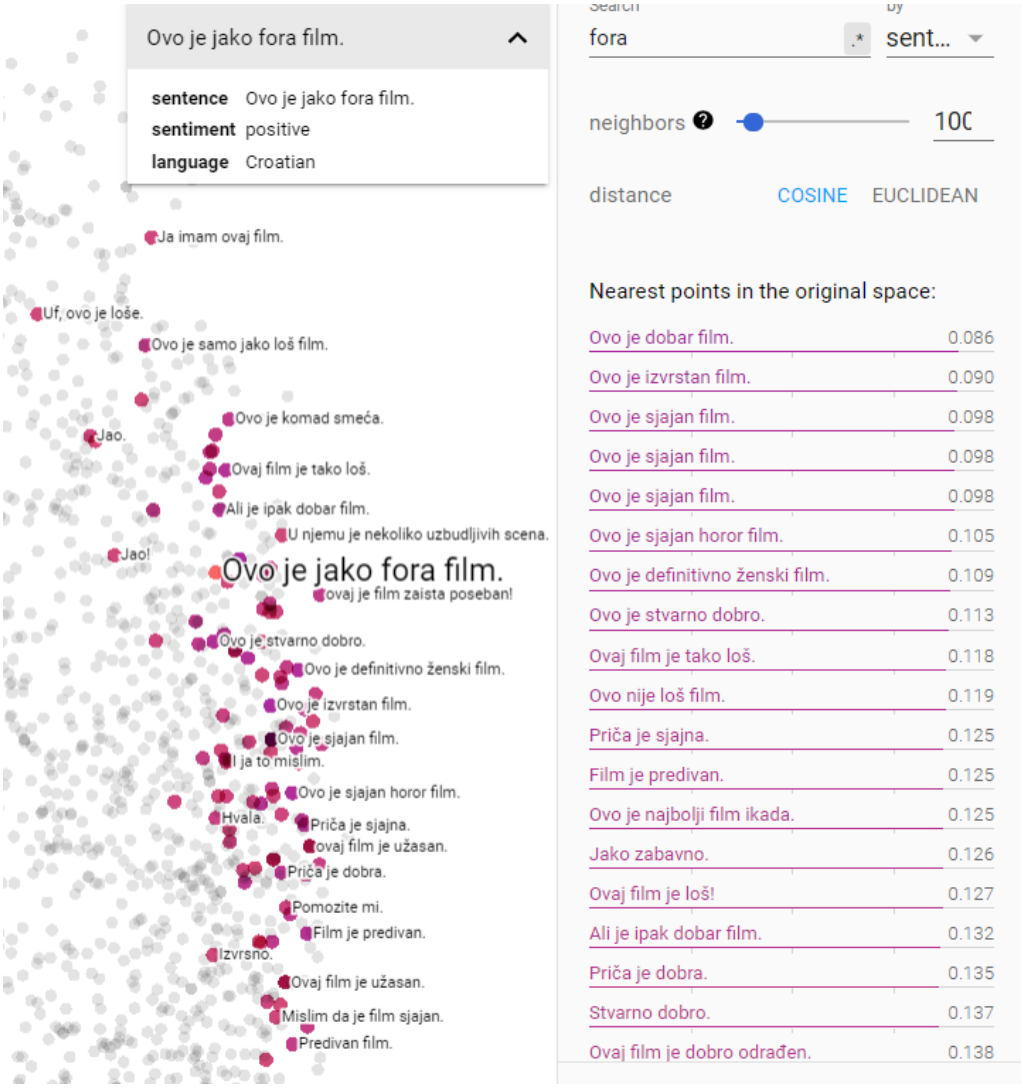
multilingual-uncased i EMBEDDIA/crosloengual-bert. Izdvojit ćemo iste primjere rečenica koje smo uzeli za ove modele kako bismo vidjeli koje rečenice model pronalazi kao najsličnije:



Slika 16 - Primjer rečenice na engleskom jeziku za model classla/bcms-bertic

Na slici 16, model za rečenicu „This is a very cool movie.“ pronalazi njoj slične rečenice po značenju. Kao i kod prijašnjih modela, BERTić pronalazi rečenice slične sintakse i značenja, no s malo boljim rezultatima. Sve od navedenih sličnih rečenica vidljivih u inspektorskom izborniku počinju s riječju „this“, kao i označena rečenica. Osim toga, model pronalazi sinonime određenih riječi koje prenose isto ili slično značenje kao i označena rečenica. Kao npr. pridjevi „great“, „good“, „funny“, „greatest“, „excellent“, „surprising“ itd. Naravno, vidljive su i rečenice negativnog sentimenta, no one sadrže sličnu sintaksu te je većina rečenica ipak isto pozitivnog sentimenta kao i označena rečenica. Od 100 najbližih izoliranih rečenica, njih 60 imaju pozitivan sentiment, 28 negativan i 12 neutralan.

Možemo vidjeti koje je rečenice model prepoznao kao najsličnije istoj ovoj rečenici prevedenoj na hrvatski jezik:



Slika 17 - Primjer rečenice na hrvatskom jeziku za model classla/bcms-bertic

Promotrimo li inspektorski izbornik na slici 17 možemo uočiti da pronađene rečenice također sadrže slične ili iste riječi, sintaksu te u većini slučajeva isti sentiment kao i označena rečenica. Što se tiče strukture i značenja, ovaj model pronalazi sličnije rečenice na hrvatskom jeziku od modela bert-base-multilingual-uncased i EMBEDDIA/crosloengual-bert. Ako izoliramo najbližih 100 rečenica, njih 60 će imati pozitivan sentiment, 27 negativan i 14 neutralan.

Model classla/bcms-bertic omogućuje jasnu vizualizaciju razlika u rečenicama s obzirom na jezik. Na temelju primjera rečenica, ovaj model, kao i modeli bert-base-multilingual-uncased i EMBEDDIA/crosloengual-bert pronalazi približno isti broj rečenica pozitivnog sentimenta među najbližim rečenicama, no za razliku od navedenih modela, BERTić pokazuje bolje rezultate s obzirom na semantičko značenje i sintaktičke konstrukcije rečenica, pogotovo u primjeru rečenice na hrvatskom jeziku.

4.5. Rezultati

MODEL	Jezici kojima je prilagođen	Jasna raspodjela po jeziku	Uvježban na analizi sentimenta	Jasna raspodjela po sentimentu
bert-base-multilingual-uncased	——	DA	NE	NE
nlptown/bert-base-multilingual-uncased-sentiment	engleski nizozemski njemački francuski španjolski talijanski	NE	DA	DA
EMBEDDIA/crosloengual-bert	hrvatski slovenski engleski	DA	NE	NE
EMBEDDIA/crosloengual-bert-sentiment	hrvatski slovenski engleski	DA	DA	DA
classla/bcms-bertic	hrvatski bosanski srpski crnogorski	DA	NE	NE

Tablica 4 - Pregled modela

Od navedenih modela u tablici 4 svi osim nlptown/bert-base-multilingual-uncased-sentiment su uspješni u raspodjeli korištenih vektorskih prikaza riječi na temelju različitih jezika. Najučinkovitiji modeli u toj raspodjeli su bert-base-multilingual-uncased i EMBEDDIA/crosloengual-bert pošto imaju najmanju količinu preklapanja između rečenica na hrvatskom i engleskom jeziku. Modeli koji su uvježbani na analizi sentimenta jasno grupiraju rečenice po značenju s obzirom na sentiment, no model EMBEDDIA/crosloengual-bert-sentiment je jedini koji vizualno odvaja rečenice i po jeziku i sentimentu. Pošto je ovaj model uvježban na analizi sentimenta te prilagođen hrvatskom i engleskom jeziku, pokazuje najbolje rezultate u vizualizaciji s obzirom na korišten skup podataka.

Kod modela uvježbavanih na analizi sentimenta proučene su rečenice koje su modeli vizualno grupirali među rečenice s drugačijim sentimentom. Prepoznato je da su modelima probleme stvarale rečenice koje u sebi sadrže više različitih konotacija. Moguće je da model ima poteškoća s određivanjem sentimenta rečenice kada se u njoj nalaze riječi pozitivnih i

negativnih konotacija. Osim toga, primijećeno je da modeli ne mogu prepoznati sarkazam koji sadrži negativan sentiment, već tumače te rečenice u doslovnom smislu.

Na temelju primjera rečenica „This is a very cool movie.“ i „Ovo ja jako fora film.“ proučene su izvedbe modela bert-base-multilingual-uncased, EMBEDDIA/crosloengual-bert i classla/bcms-bertic fokusirajući se na njihovu sposobnost pronalaženja njima najsličnijih rečenica po značenju. Sva tri modela su od najsličnijih 100 rečenica pronašla većinom one koje sadrže isti – pozitivan – sentiment kao i označene rečenice. Koristeći primjer rečenice na engleskom jeziku, modeli EMBEDDIA/crosloengual-bert i classla/bcms-bertic su pronašli njoj semantički i sintaktički sličnije rečenice od modela bert-base-multilingual-uncased. Na temelju rečenice na hrvatskom jeziku, classla/bcms-bertic model je pokazao znatno bolje rezultate što je za očekivati s obzirom na činjenicu da je prilagođen hrvatskom i njemu slična 3 jezika, dok je bert-base-multilingual-uncased uvježbavan samo nenadziranim učenjem na maloj količini hrvatskih tekstova. CroSloEngual BERT je uvježbavan na sličnom broju podataka na hrvatskom jeziku kao i BERTić, no bez podataka o ostalim jezicima.

5. Zaključak

Sa sve većom primjenom sustava strojnog učenja u različitim aspektima ljudskog života, znanstvenicima i istraživačima postaje sve važnije analizirati kako modeli interpretiraju podatke. Ovaj proces dodatno olakšavaju javno dostupne besplatne tehnologije OPJ-a koje su korištene u sklopu projektnog dijela ovog rada. HuggingFace biblioteka modela transformera omogućuje korištenje već prethodno uvježbanih modela prilagođenih različitim jezicima i zadacima OPJ-a. Zahvaljujući radnom okviru Tensorflow Embedding Projector-u moguće je na intuitivan i interaktivan način vizualizirati te proučavati vektorske prikaze riječi i njihove međusobne odnose. Demonstracijom izvedbe modela u vizualizaciji korištenog skupa podataka istaknute su neke od sličnosti i razlika u njihovoj mogućnosti izvršavanja traženih zadataka. S obzirom na skup podataka koji sadržava neformalne rečenične strukture i izraze te uzimajući u obzir činjenicu kojim su jezicima i zadacima OPJ-a modeli prilagođeni, rezultati su se pokazali vrlo zadovoljavajućima. Ovisno o tome na koji aspekt skupa podataka se želimo fokusirati, moguće je izabrati model koji će pružiti najbolju vizualizaciju traženog aspekta i analizu njegova načina razumijevanja podataka.

S dostupnom količinom digitalnih podataka koji se neprestano povećavaju i pojavom sve sofisticiranijih i preciznijih algoritama, popularnost OPJ-a će zasigurno nastaviti rasti. Korištene svakom dostupne tehnologije OPJ-a te biblioteke skupova podataka na različitim jezicima otvaraju mogućnost lakšeg stvaranja, korištenja, uvježbavanja i prilagođavanja modela transformera čime se otvara prostor za nova istraživanja u kontekstu OPJ-a, strojnog učenja i umjetne inteligencije.

6. Literatura

Ahmad, P. & Dang, S. (2014). A Comparative Study on Text mining Techniques. *International Journal of Science and Research.* Volume 2. 2222-2226. Preuzeto s https://www.researchgate.net/publication/270704468_A_Comparative_Study_on_Text_mining_Techniques [4. srpnja 2021.]

Allen, F. J. (2003). Natural language processing. *Encyclopedia of Computer Science*. John Wiley and Sons Ltd., 1218–1222. https://dl.acm.org/doi/10.5555/1074100.1074630 [4. srpnja 2021.]

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, *35*(8), 1798-1828. https://doi.org/10.1109/tpami.2013.50

Bengio, Y.. (2012). Deep Learning of Representations for Unsupervised and Transfer Learning. *Proceedings of ICML Workshop on Unsupervised and Transfer Learning, in Proceedings of Machine Learning Research*, 27:17-36 Preuzeto s http://proceedings.mlr.press/v27/bengio12a.html [4. srpnja 2021.]

Calijorne Soares, M., & Parreiras, F. (2020). A literature review on question answering techniques, paradigms and systems. *Journal Of King Saud University - Computer And Information Sciences*, *32*(6), 635-646. https://doi.org/10.1016/j.jksuci.2018.08.005 [4. srpnja 2021.]

Clark, K., & Luong, T. (2020). More Efficient NLP Model Pre-training with ELECTRA. Preuzeto s from https://ai.googleblog.com/2020/03/more-efficient-nlp-model-pre-training.html [4. srpnja 2021.]

Devlin, J., & Chang, M. (2018). *Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing*. Google AI Blog. Preuzeto s https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html [4. srpnja 2021.]

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings Of The 2019 Conference Of The North*, 4171–4186. https://doi.org/10.18653/v1/n19-1423

Dialani, P. (2020). *NLP v/s NLU v/s NLG*. Analyticsinsight.net. Preuzeto s https://www.analyticsinsight.net/nlp-vs-nlu-vs-nlg/ [4. srpnja 2021.]

Donges, N. (2021). *A Guide to RNN: Understanding Recurrent Neural Networks and LSTM*. Built In. Preuzeto s https://builtin.com/data-science/recurrent-neural-networks-and-lstm [4. srpnja 2021.]

Google Colaboratory. (2021). *Google Colaboratory*. Colab.research.google.com. Preuzeto s https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index [4. srpnja 2021.]

HuggingFace.co. *Hugging Face – The AI community building the future.*. Huggingface.co. (2021). Preuzeto s https://huggingface.co/ [4. srpnja 2021.]

Hussein, D. (2018). A survey on sentiment analysis challenges. *Journal Of King Saud University - Engineering Sciences*, *30*(4), 330-338. https://doi.org/10.1016/j.jksues.2016.04.002 [4. srpnja 2021.]

IBM. (2020). *What are Recurrent Neural Networks?*. Ibm.com. Preuzeto s https://www.ibm.com/cloud/learn/recurrent-neural-networks [4. srpnja 2021.]

Jurafsky, D., & Martin, J. (2009). *Speech and Language Processing* (2nd ed.). Prentice Hall. https://web.stanford.edu/~jurafsky/slp3/ [4. srpnja 2021.]

Kaur, J. (2021). *What are the Differences Between NLP, NLU, and NLG?*. Xenonstack.com. Preuzeto s https://www.xenonstack.com/blog/difference-between-nlp-nlu-nlg [4. srpnja 2021.]

Kumar, C. (2018). *NLP vs NLU vs NLG (Know what you are trying to achieve) NLP engine (Part-1).* Towards Data Science. Preuzeto s https://towardsdatascience.com/nlp-vs-nlu-vs-nlg-know-what-you-are-trying-to-achieve-nlp-engine-part-1-1487a2c8b696 [4. srpnja 2021.]

Lanners, Q. (2019). *Neural Machine Translation*. Towards Data Science. Preuzeto s https://towardsdatascience.com/neural-machine-translation-15ecf6b0b [4. srpnja 2021.]

Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science* (2nd ed.). New York. Marcel Decker, Inc.

Liu, Z., Lin, Y., & Sun, M. (2020). Representation Learning and NLP. *Representation Learning For Natural Language Processing*, 1-11. https://doi.org/10.1007/978-981-15-5573-2_1

Ljubešić, N. & Lauc, D. (2021). BERTić -- The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *CoRR*. Preuzeto s https://arxiv.org/abs/2104.09243 [4. srpnja 2021.]

Naushad, R. (2021). *Tensorboard Embedding Projector—Visualizing High Dimensional Vectors with t-SNE or PCA.*. Medium. Preuzeto s https://medium.com/analytics-vidhya/tensorboard-embedding-projector-visualizing-high-dimensional-vectors-with-t-sne-or-pca-d616e222247a [4. srpnja 2021.]

Nayak, P. (2019). *Understanding searches better than ever before*. Google. Preuzeto s https://blog.google/products/search/search-language-understanding-bert/ [4. srpnja 2021.]

Nikulski, J. (2021). *How to Use Transformer-based NLP Models*. Towards Data Science. Preuzeto s https://towardsdatascience.com/how-to-use-transformer-based-nlp-models-a42adbc292e5 [4. srpnja 2021.]

Pandey, P. (2019). *Visualizing Bias in Data using Embedding Projector*. Towards Data Science. Preuzeto s https://towardsdatascience.com/visualizing-bias-in-data-using-embedding-projector-649bc65e7487 [4. srpnja 2021.]

Rocca, J. (2019). *Introduction to Markov chains*. Towards Data Science. Preuzeto s https://towardsdatascience.com/brief-introduction-to-markov-chains-2c8cab9c98ab [4. srpnja 2021.]

Schiappa, M. (2021). *Popular Downstream Tasks for Video Representation Learning*. Towards Data Science. Preuzeto s https://towardsdatascience.com/popular-downstream-tasks-for-video-representation-learning-8edbd8dc19c1 [4. srpnja 2021.]

Skansi, S., & Lauc, D. (2018). Analogical Reasoning and Word-Meanings in a Multidimensional Space. *Filozofska Istraživanja*, *38*(1), 5-16. https://doi.org/10.21464/fi38101

Smilkov, D., Thorat, N., Nicholson, C., Reif, E., Viégas, F.B., & Wattenberg, M. (2016). Embedding Projector: Interactive Visualization and Interpretation of Embeddings. Preuzeto s https://arxiv.org/abs/1611.05469 [4. srpnja 2021.]

Soni, D. (2018). *Introduction to Markov Chains*. Towards Data Science. Preuzeto s https://towardsdatascience.com/introduction-to-markov-chains-50da3645a50d#_=_ [4. srpnja 2021.]

Šuman, S. (2021). Pregled metoda obrade prirodnih jezika i strojnog prevođenja. *Zbornik Veleučilišta u Rijeci, 9* (1), 371-384. Preuzeto s https://hrcak.srce.hr/257657 [4. srpnja 2021.]

Thakkar, G. (2021). Multi-task Learning for Cross-Lingual Sentiment Analysis. CLEOPATRA Workshop 2021. https://doi.org/10.5446/52943

TensorFlow.org. (2021). TensorFlow. Preuzeto s https://www.tensorflow.org/ [4. srpnja 2021.]

Theodoridis, S. (2020). Neural Networks and Deep Learning. *Machine Learning*, 901-1038. https://doi.org/10.1016/b978-0-12-818803-3.00030-1

Tyagi, P. (2021). *Markov chain : Mathematical formulation, Intuitive Explanation & Applications*. Analytics Vidhya. Preuzeto s https://www.analyticsvidhya.com/blog/2021/02/markov-chain-mathematical-formulation-intuitive-explanation-applications/ [4. srpnja 2021.]

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. *CoRR*. Preuzeto s https://arxiv.org/abs/2006.07890 [4. srpnja 2021].

Uszkoreit, J. (2017*). Transformer: A Novel Neural Network Architecture for Language Understanding*. Google AI Blog. Preuzeto s https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html [4. srpnja 2021.]

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. et al. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (str. 6000–6010). https://dl.acm.org/doi/10.5555/3295222.3295349.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168 [4. srpnja 2021.]

7. Popis slika

8. Popis tablica

9. Sažetak

Obrada prirodnog jezika postaje sve aktivnijim područjem istraživanja. Razvoj tehnika reprezentativnog učenja od lokalnih do distribuiranih reprezentacija omogućio je tehnologijama obrade prirodnog jezika bolje razumijevanje semantičkog značenja i konteksta riječi i rečenica na prirodnim jezicima. Veliki napredak ostvaren je dvosmjernim treniranjem modela transformera čija je izvedba prikazana u ovom radu vizualizacijom skupa podataka IMDB recenzija na hrvatskom i engleskom jeziku.

Ključne riječi: obrada prirodnog jezika, reprezentativno učenje, lokalna reprezentacija, distribuirana reprezentacija, transformeri, BERT

10. Summary

Natural language processing has become an increasingly active area of research. The development of representative learning techniques from local to distributed representations has enabled natural language processing technology to better understand the semantic meaning and context of words and sentences in natural languages. Great progress has been made in the bidirectional training of transformer models whose performance is presented in this paper by their visualization of a set of IMDB reviews in Croatian and English language.

Keywords: natural language processing, representation learning, local representation, distributed representation, transformers, BERT