Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

Rino Čala, Ivan Martinović

Implementacija konverzacijskih modela s različitim crtama osobnosti

Zagreb, 2022.

Ovaj rad izrađen je na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave pod vodstvom prof. dr. sc. Jana Šnajdera i mag. ing. Martina Tuteka i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2021./2022.

Sadržaj

1 Uvod

U potrazi za nekom uslugom na internetu svakodnevno nailazimo na konverzacijske modele (engl. *chatbot*), računalne programe namijenjene automatskom odgovaranju na korisničke upite. Takvi računalni programi najčešće su ograničeni predefiniranim naredbama, a uloga korisnika zapravo je praćenje uputa koje pred njih stavlja računalni program s ciljem dobivanja određene informacije ili usluge. Takvi konverzacijski modeli usko su povezani uz određenu domenu (npr., banke, telekomi, internetske trgovine), što ih čini razmjerno ograničenima. Budući da su različita stanja u kojima se računalni program može naći unaprijed određena, ova vrsta konverzacijskih modela u praksi se često implementira kao jednostavan automat s konačnim brojem stanja. S druge strane, postoje konverzacijski modeli otvorenog tipa čije ponašanje nije unaprijed određeno. Takav računalni program zapravo nastoji generirati odgovore na korisničke upite što sličnije odgovorima koje bi generirao čovjek. Ako želimo da računalo bude sposobno voditi razgovor kao čovjek, jasno nam je da računalo mora razumjeti i obrađivati prirodni jezik te izvoditi zaključke. Osim toga, ono mora moći generirati neograničen skup odgovora koji uzimaju u obzir prošlost razgovora i sugovornikova psihološka stanja te moraju biti sukladna s općim znanjem.

Rješavanjem ovakvih, računalu izrazito teških, a čovjeku jednostavnih problema, bavi se područje strojnog učenja (engl. *machine learning*). Strojno učenje programiranje je računala s ciljem optimizacije nekog kriterija uspješnosti temeljem podatkovnih primjera ili prethodnog iskustva. Raspolažemo modelom koji je određen do na neke parametre, a učenje modela svodi se na optimizaciju tih parametara koristeći podatkovne primjere ili prethodno iskustvo [1]. Duboko učenje (engl. *deep learning*) područje je strojnog učenja inspirirano modelima sličnim ljudskom mozgu te bi imao sposobnost imitirati funkciju ljudskog mozga. Ako bismo imali takav model, mogli bismo riješiti algoritamski teške probleme koje čovjek rješava bez previše razmišljanja, a jedan od takvih problema svakako je i generiranje rečenica, odnosno sudjelovanje u razgovoru.

Cilj ovog rada razvijanje je konverzacijskog modela sposobnog za sudjelovanje u razgovoru s čovjekom. Svaki čovjek posjeduje osobnost koja se očituje njegovim oblicima ponašanja, psihološkim stanjima i interakcijama s okolinom. Budući da konverzacijski model mora voditi razgovor poput čovjeka, trebao bi posjedovati i osobnost koja ga karakterizira. Potaknuti time, u okviru ovoga rada bavimo se razvojem konverzacijskih modela koji će, uz mogućnost konverzacije s čovjekom, posjedovati

i crte osobnosti prema MBTI (engl. *Myers-Briggs Type Indicator*) modelu osobnosti. Koliko nam je poznato, ovaj rad je prvi rad koji za cilj ima povezivanje psiholoških karakteristika osobnosti s konverzacijskim modelima.

Na početku rada predstavili smo područje strojnog učenja te duboko učenje kao njegovu zasebnu granu. Nakon toga predstavljeni su tradicionalni pristupi u području obrade prirodnog jezika (engl. *natural language processing*), a potom su prikazane novije arhitekture korištene u okviru rada. U trećem poglavlju opisani su najčešće korišteni upitnici osobnosti, ukratko je predstavljeno područje procjene osobnosti te je opisan skup podataka PANDORA [2]. Zatim su u četvrtom i najvažnijem poglavlju prikazani svi dijelovi našeg razvijenog sustava. Prvo je opisan postupak pripreme skupa podataka za učenje modela zaduženog za klasifikaciju crta osobnosti iz teksta, a nakon toga opisan je i sam postupak učenja tog modela. Zatim je pojašnjen postupak stvaranja i filtriranja novog, velikog skupa podataka koji smo izradili sa svrhom učenja konverzacijskog modela. Stvaranje tog skupa podataka uključivalo je preuzimanje oko 81 milijuna dijaloga s *Reddita*, što je trajalo nekoliko mjeseci. Potom je opisan i sam postupak učenja konverzacijskih modela s različitim crtama osobnosti, a prikazan je i postupak neformalne evaluacije modela pomoću mrežno dostupnog MBTI upitnika osobnosti. Naposljetku, prikazana je web-aplikacija izrađena u svrhu demonstracije rada naših modela.

2 Metoda

2.1 Strojno učenje

Strojno učenje postupak je programiranja računala s ciljem optimizacije nekog kriterija uspješnosti na temelju podatkovnih primjera. Tri su osnovne komponente svakog algoritma strojnog učenja: (1) model, (2) funkcija gubitka te (3) optimizacijski postupak. Također, postoje tri osnovna pristupa strojnom učenju: (1) nadzirano, (2) nenadzirano i (3) podržano učenje. Nadzirano učenje podrazumijeva da raspolažemo skupom označenih primjera, odnosno za svaki ulazni primjer $\mathbf{x}$ postoji odgovarajuća oznaka y.

Promotrimo tri navedene osnovne komponente strojnog učenja kroz prizmu nadziranog učenja. Cilj nadziranog učenja jest naučiti preslikavanje ulaza $\mathbf{x}$ na izlaz y pomoću skupa uzoraka za učenje (engl. *training set*):

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(i)}, y^{(i)} \right) \right\}_{i=1}^{N} \subseteq \mathcal{X} \times \mathcal{Y},$$

pri čemu je N veličina skupa za učenje, $\mathcal{X}$ prostor primjera, a $\mathcal{Y}$ prostor oznaka. Prostor primjera $\mathcal{X}$ uglavnom je višedimenzijski vektorski prostor, a svaka dimenzija tog prostora predstavlja neku diskretnu ili kontinuiranu značajku ulaznog primjera. U slučaju da rješavamo klasifikacijski problem, npr., na osnovu objava korisnika na nekom forumu želimo klasificirati njegov ili njezin tip osobnosti u jednu od predefiniranih kategorija osobnosti, prostor oznaka bili bi mogući tipovi osobnosti kojima neka osoba može pripadati. U slučaju regresijskog problema, prostor oznaka je kontinuiran (npr., želimo procijeniti u kojoj mjeri neka osoba pripada svakom od unaprijed definiranih tipova osobnosti).

Pretpostavimo da postoji funkcija h^* koja savršeno preslikava ulaze iz prostora primjera $\mathcal{X}$ u prostor oznaka $\mathcal{Y}$. Budući da je funkcija h^* upravo ono što nas zanima, definiramo funkciju $h : \mathcal{X} \to \mathcal{Y}$ kojom ćemo, pomoću skupa za učenje $\mathcal{D}$, pokušati što bolje aproksimirati nepoznatu funkciju h^*. Funkcija h bit će definirana do na parametre $\boldsymbol{\theta}$, koje za vrijeme optimizacije prilagođavamo. Ovako definiranu funkciju h nazivamo hipotezom [3]. Izgledno je kako će za različite vrijednosti parametara $\boldsymbol{\theta}$ hipoteza h različito preslikavati prostor primjera u prostor oznaka, a skup svih mogućih preslikavanja koje dobivamo variranjem parametara $\boldsymbol{\theta}$ nazivamo modelom:

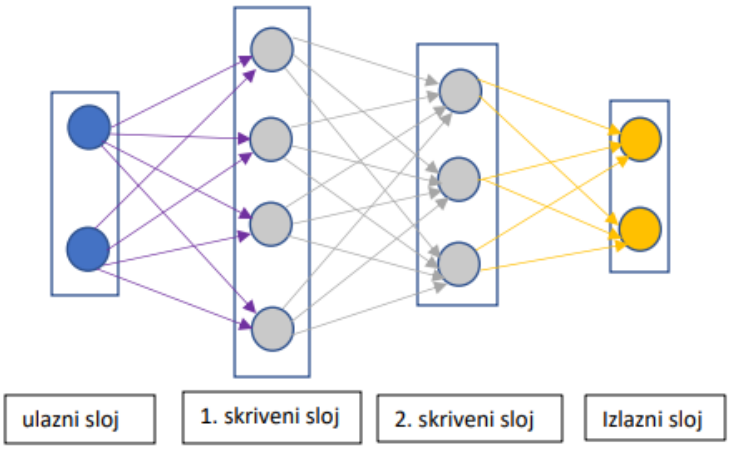
$$\mathcal{H} = \{h(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta}}\,.$$

Nakon što smo definirali kako izgledaju naše hipoteze, potrebno je odrediti neki kriterij prema kojemu ćemo odlučiti koja je najbolja hipoteza iz skupa svih hipoteza.

Definiramo funkciju gubitka $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, koja nam govori koliko se izlaz naše hipoteze za i-ti podatkovni primjer $h\left(\mathbf{x}^{(\mathbf{i})}; \boldsymbol{\theta}\right)$ razlikuje od točne oznake istog primjera $y^{(i)}$. Naš je cilj minimizirati tu razliku, odnosno minimizirati iznos funkcije gubitka L za sve primjere. Tako ćemo pronaći najbolju aproksimaciju naše savršene funkcije h^*. Budući da su hipoteze u parametrima često nelinearne funkcije, minimizaciju funkcije gubitka, odnosno funkcije pogreške ako u obzir uzmemo sve primjere iz skupa za učenje, nećemo moći provesti analitički. Zbog toga kao optimizacijski postupak koristimo neki iterativan postupak, kao što je naprimjer, postupak gradijentnog spusta.

2.1.1 Duboko učenje

Područje dubokog učenja formira hipotezu h kao kompozitnu strukturu koju čini više jednostavnih procesnih elemenata koji obavljaju jednostavne računske operacije, a po analogiji s ljudskim mozgom, te procesne elemente u području dubokog učenja nazivamo neuronima. Neuroni su međusobno povezani težinama čiji iznosi predstavljaju parametre hipoteze $\boldsymbol{\theta}$, koje nastojimo prilagoditi za vrijeme optimizacijskog postupka. Više neurona organiziranih u slojeve koji su međusobno povezani težinama nazivamo umjetnom neuronskom mrežom. Ako neuroni u pojedinom sloju nisu međusobno povezani, govorimo o unaprijednoj umjetnoj neuronskoj mreži. Svaki sloj unaprijedne umjetne neuronske mreže predstavlja jednu funkciju, a čitavu mrežu možemo shvatiti kao kompoziciju tih funkcija.



Slika 1: Primjer unaprijedne umjetne neuronske mreže s ulaznim slojem od 2 neurona, 2 skrivena sloja i izlaznim slojem s 2 neurona. Slika preuzeta iz [4].

Težine sa Slike 1 možemo grupirati u matricu težina koju najčešće označavamo s $\mathbf{W}_i$, pri čemu i označava i-ti sloj. Ovakva unaprijedna umjetna neuronska mreža obavlja

preslikavanje $h : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. Formalno, neuronsku mrežu prikazanu na slici možemo predstaviti kao funkciju h, gdje $h(\mathbf{x}) = f_3\left(\mathbf{W}_3 \cdot f_2(\mathbf{W}_2 \cdot f_1\left(\mathbf{W}_1 \cdot \mathbf{x}\right))\right)$, pri čemu su f_i aktivacijske funkcije koje primjenjuju nelinearnu transformaciju na svaku dimenziju vektora (engl. *element-wise*), dok množenje predstavlja matrično množenje.

2.2 Obrada prirodnog jezika

Obrada prirodnog jezika (engl. *natural language processing*, NLP) područje je umjetne inteligencije koje za cilj ima omogućiti računalima razumijevanje jezika na sličan način na koji ga razumiju ljudi. Razumijevanje jezika zahtjevan je zadatak za računalo. Do pojave strojnog učenja, računalni sustavi za razumijevanje i obradu prirodnog jezika uglavnom su bili sustavi temeljeni na pravilima (engl. *rule-based systems*). Pojava dubokog učenja omogućila je učinkovito rješavanje velikog broja zadataka iz područja obrade prirodnog jezika, a neki od tih zadataka su: označavanje vrste riječi (engl. *part of speech tagging*), prepoznavanje vlastitih imena (engl. *named entity recognition*), raspoznavanje govora (engl. *speech recognition*), generiranje teksta (engl. *natural language generation*) i dr. U okviru ovog rada bavili smo se zadacima klasifikacije crta osobnosti korisnika na temelju njegovih objava na forumu *Reddit*, a nakon toga i zadatkom generiranja teksta na engleskom jeziku.

2.2.1 Vektorske reprezentacije riječi

Kako bi računalni program bio u stanju razumjeti jezik, jasno nam je kako ćemo jezik, odnosno njegove sastavnice – riječi – morati prikazati u obliku pogodnom za računalo. Budući da algoritmi strojnog učenja preferiraju numeričke reprezentacije, riječi ćemo prikazati kao vektore brojčanih vrijednosti. Tri su osnovna načina prikazivanja riječi kao diskretnih objekata u obliku višedimenzionalnih vektora značajki:

- jednojedinične reprezentacije riječi (engl. *one-hot representations*);
- vektori zasnovani na frekvenciji pojavljivanja riječi u korpusu (engl. *distributional count-based vectors*);
- distribucijske reprezentacije riječi (engl. *distributed representations*), poznatije pod imenom *Word-embeddings*.

Danas se najčešće koriste distribucijske reprezentacije riječi (engl. *word-embeddings*) naučene pomoću jednostavnog modela dubokog učenja – *Word2Vec* vektorske reprezentacije riječi [5]. Takve reprezentacije riječi temelje se na distribucijskoj pretpostavci (engl. *distributional hypothesis*), koja pretpostavlja kako riječi koje se pojavljuju

u sličnom kontekstu imaju slično značenje [6]. Ideja je da model korišten za izgradnju vektorskih reprezentacija riječi na osnovu okolnih riječi koje predstavljaju kontekst predvidi riječ koja nedostaje (engl. *CBOW model*) ili na osnovu jedne riječi predvidi kontekst te riječi (engl. *skip-gram model*). Posebna pogodnost ovakvih modela je što ne zahtijevaju ručno označene podatke, odnosno vrlo jednostavno je automatski izgraditi skup primjera za učenje. Danas su dostupne prednaučene vektorske reprezentacije riječi nastale učenjem na velikoj količini nestrukturiranog teksta dostupnog na internetu, npr., tekst s mrežnog sjedišta *Wikipedia*.[1]

2.2.2 Povratna neuronska mreža

Povratne neuronske mreže (engl. *Recurrent neural networks*, RNN) namijenjene su obradi slijednih podataka (engl. *sequence data*) vodeći pritom računa o poretku tih podataka unutar slijeda. Rečenice prirodnog jezika svakako su primjer slijednih podataka gdje su nam povratne neuronske mreže korisne. Ideja iza povratnih neuronskih mreža pamćenje je informacija iz prošlosti, što pomoću unaprijedne umjetne neuronske mreže s fiksnom duljinom ulaza nije lako postići. Upravo je mogućnost obrade podataka s varijabilnom duljinom najvažnija značajka povratnih neuronskih mreža. Karakteristika povratnih neuronskih mreža postojanje je povratne veze kojom naglašavamo kako trenutačni izlaz, koji nazivamo skrivenim stanjem (engl. *hidden state*), ovisi o prethodnom skrivenom stanju i trenutačnom ulazu.

Izraz za ažuriranje skrivenog stanja jednostavne povratne ćelije (engl. *vanilla recurrent neural network cell*) jest:

$$\mathbf{h}_t = f\left(\mathbf{h}_{t-1}, \mathbf{x}_t\right) = f\left(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{xh}\mathbf{x}_t + \mathbf{b}_h\right).$$

Pritom su $\mathbf{W}_{hh}$ i $\mathbf{W}_{xh}$ matrice parametara koje se dijele za svaki vremenski trenutak t, $\mathbf{h}_t$ predstavlja vektor skrivenog stanja u trenutku t, vektor $\mathbf{x}_t$ predstavlja trenutni ulaz (npr., vektorska reprezentacija trenutne riječi u rečenici), dok je $\mathbf{b}_h$ vektor pomaka. Postupak optimizacije povratnih neuronskih mreže postupak je propagiranja pogreške unatrag kroz vrijeme (engl. *backpropagation through time*) temeljen na već spomenutom algoritmu gradijentnog spusta. Zbog problema jednostavne povratne neuronske mreže, kao što su npr., problem eksplodirajućih i nestajućih gradijenata (engl. *exploding and vanishing gradients problem*), razvijene su naprednije arhitekture ćelija povratnih neuronskih mreža, od kojih su najvažnije: ćelija s dugoročnim pamćenjem (engl. *long short-term memory cell*, LSTM) [7] te propusna povratna ćelija (engl. *gated recurrent unit*, GRU) [8].

[1] `https://www.wikipedia.org/`

2.3 Modeli temeljeni na slojevima pozornosti

Dugi niz godina povratne neuronske mreže u kombinaciji s prednaučenim vektorskim reprezentacijama riječi ostvarivale su najbolje rezultate na velikom broju zadataka iz područja obrade prirodnog jezika. Međutim, povratne neuronske mreže imaju svojih nedostataka, a jedan od najvećih svakako je slijedno računanje skrivenih stanja kroz vremenske trenutke. Slijedni izračun onemogućava paralelizaciju izračuna skrivenih stanja te stvara probleme u slučaju da raspolažemo dugim ulaznim tekstovima. Osim toga, pogledamo li kao primjer zadatak strojnog prevođenja (engl. *machine translation*), jasno je kako će standardne koder-dekoder arhitekture [9, 10] koje čitavu ulaznu rečenicu kodiraju u vektor fiksne dimenzije, imati probleme s dugim ulaznim rečenicama jer postoji mogućnost kako će neke važne informacije iz prethodnih skrivenih stanja u izlaznoj reprezentaciji biti zaboravljene. U spomenutoj arhitekturi koder radi tako da ulaznu rečenicu $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$ transformira u izlazni vektor $\mathbf{c}$, koji sadrži reprezentaciju cijele ulazne rečenice. Obično se kao koder koristi povratna neuronska mreža te vrijede izrazi [11]:

$$\mathbf{h}_t = f\left(\mathbf{h}_{t-1}, \mathbf{x}_t\right), \quad \mathbf{c} = g\left(\{\mathbf{h}_1, \cdots, \mathbf{h}_T\}\right),$$

pri čemu je $\mathbf{h}_t$ skriveno stanje u trenutku t, vektor $\mathbf{c}$ vektor je nastao transformacijom niza skrivenih stanja i nazivamo ga kontekstom (engl. *context vector*), dok su f i g nelinearne funkcije. U radu [10] koder je povratna ćelija s dugoročnim pamćenjem (engl. *LSTM cell*), dok za funkciju g vrijedi: $g\left(\{\mathbf{h}_1, \cdots, \mathbf{h}_T\}\right) = \mathbf{h}_T$. S druge strane, dekoder učimo predviđati sljedeću riječ y_t na temelju vektora konteksta $\mathbf{c}$ i svih prethodno generiranih riječi $\{y_1, \cdots, y_{t-1}\}$. Zadaća je dekodera maksimizirati vjerojatnost izlaznog slijeda $y = (y_1, \cdots, y_{T'})$:

$$p\left(y\right) = \prod_{t=1}^{T'} p\left(y_t \mid \{y_1, \cdots, y_{t-1}\}, \mathbf{c}\right). \tag{1}$$

Uvjetnu vjerojatnost iz izraza (1) možemo modelirati povratnom neuronskom mrežom:

$$p\left(y_t \mid \{y_1, \cdots, y_{t-1}\}\right) = q\left(y_{t-1}, \mathbf{s}_t, \mathbf{c}\right), \tag{2}$$

pri čemu je q neka nelinearna funkcija, a $\mathbf{s}_t$ skriveno je stanje dekodera u vremenskom trenutku t. Važno je uočiti kako se duljina ulaznog slijeda T može razlikovati od duljine izlaznog slijeda T' jer su ove arhitekture razvijane za zadatak strojnog prevođenja, a jasno je kako rečenice s istim značenjem u različitim jezicima često imaju različitu duljinu. Ovakav formalizam karakterističan je za modele koji nastoje previdjeti sljedeću riječ (engl. *auto-regressive language model*) [12]. Primjećujemo kako generiranje

sljedeće riječi uvjetujemo uvijek jednakim vektorom konteksta $\mathbf{c}$ koji predstavlja sažetu reprezentaciju ulazne rečenice. Takav fiksni vektor konteksta usko je grlo ovakve arhitekture.

Kao rješenje tog problema autori članka [11] predstavili su mehanizam pozornosti (engl. *attention mechanism*). Za razliku od opisane koder-dekoder arhitekture, uvođenjem sloja pozornosti vjerojatnost izlaza y_t uvjetujemo uvijek novim vektorom konteksta $\mathbf{c}_t$. Izraz (2) postaje:

$$p\left(y_t \mid \{y_1, \cdots, y_{t-1}\}\right) = q\left(y_{t-1}, \mathbf{s}_t, \mathbf{c}_t\right), \tag{3}$$

pri čemu je $\mathbf{s}_t$ skriveno stanje povratne mreže dekodera u trenutku t koje računamo prema izrazu:

$$\mathbf{s}_t = f\left(y_{t-1}, \mathbf{s}_{t-1}, \mathbf{c}_t\right).$$

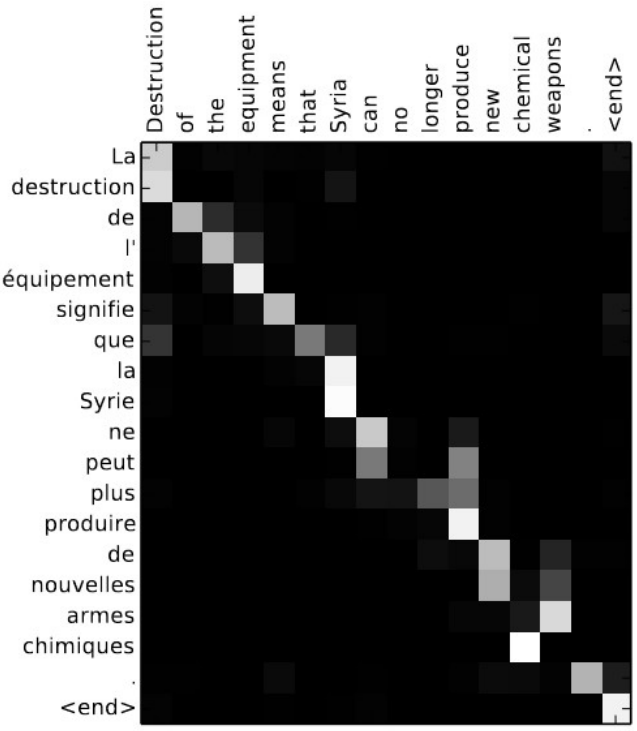
Važno je napomenuti kako autori članka [11] kao koder koriste dvosmjernu jednostavnu povratnu neuronsku mrežu (engl. *bidirectional recurrent neural network*) kao koder, što zapravo znači da jedan sloj čine dvije jednostavne povratne neuronske mreže koje u različitim smjerovima (od početka prema kraju te od kraja prema početku) kodiraju ulaznu rečenicu te proizvode skrivena stanja. Zbog toga svako skriveno stanje $\mathbf{h}_t$ ima informaciju o cijeloj ulaznoj rečenici s naglaskom na svoju bližu okolinu u oba smjera. Vektor konteksta $\mathbf{c}_t$ računa se kao težinski prosjek skrivenih stanja kodera [11]:

$$\mathbf{c}_t = \sum_{j=1}^{T} \alpha_{tj} \mathbf{h}_j,$$

pri čemu težine računamo prema izrazu:

$$\alpha_{tj} = \frac{exp\left(e_{tj}\right)}{\sum_{k=1}^{T} exp\left(e_{tk}\right)}, \quad e_{tj} = a\left(\mathbf{s}_{t-1}, \mathbf{h}_j\right). \tag{4}$$

U izrazu (4) funkciju a modeliramo potpuno povezanim slojem unaprijedne neuronske mreže, dok nam težina α_{tj} govori koliku pozornost obraćamo na skriveno stanje na poziciji j pri generiranju sljedeće riječi za vremenski trenutak t. Ako govorimo o zadatku strojnog prevođenja, težina α_{tj} govorila bi nam koliku pozornost obraćamo na riječ ulazne rečenice na poziciji j pri prevođenju riječi u vremenskom trenutku t. Upravo po toj interpretaciji govorimo o uvođenju slojeva pozornosti (engl. *attention layer*). Iznose tih težina na primjeru možemo vidjeti na Slici 2.

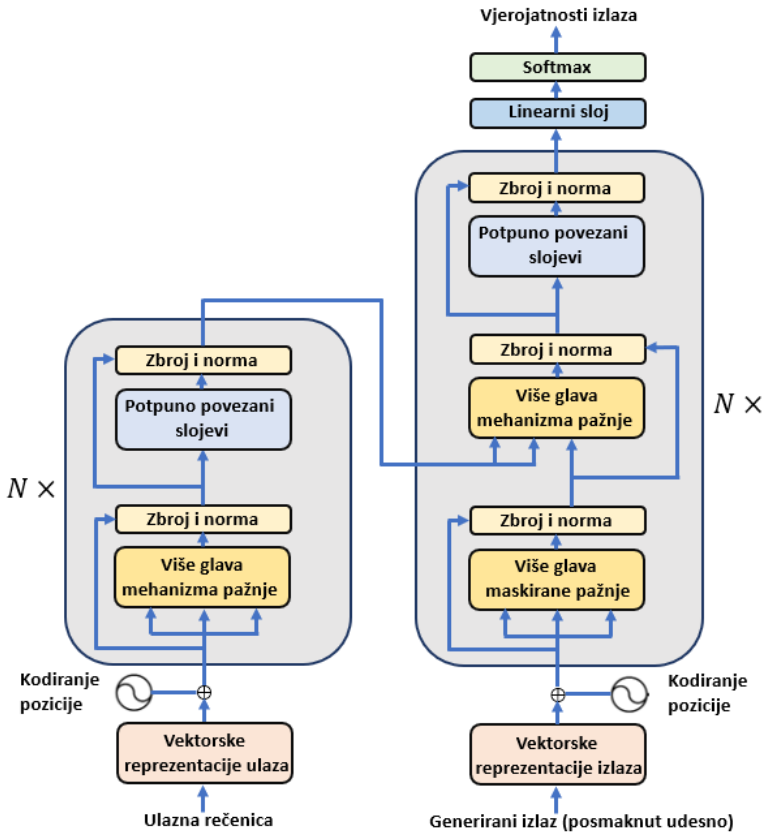


Slika 2: Iznosi težina α_{tj} na primjeru prevođenja rečenice s engleskog na francuski jezik. Na x-osi nalazi se rečenica na engleskom jeziku, dok se na y-osi nalazi prijevod te rečenice na francuski jezik. Bijeli kvadratići predstavljaju težine iznosa bliskih 1, dok crni kvadratići predstavljaju težine iznosa bliskih 0. Slika preuzeta iz članka [11].

2.3.1 *Transformer* arhitektura

Slijedno računanje skrivenih stanja povratnih neuronskih mreža bilo je problem sve do pojave nove arhitekture zvane *Transformer* [13] koja se u potpunosti temelji na ranije navedenom mehanizmu slojeva pozornosti. Ova arhitektura također je primjer koder-dekoder arhitekture. Koder na ulazu dobiva vektorske reprezentacije ulazne rečenice, a na izlazu daje kodiranu reprezentaciju te rečenice nastalu nizom nelinearnih transformacija. Slično kao i prije, zadaća je dekodera pomoću reprezentacije dobivene na izlazu kodera i riječi iz prošlih vremenskih trenutaka generirati riječ za sljedeći vremenski trenutak. Na Slici 3 prikazana je *Transformer* arhitektura, s vizualno razdvojenim koderom (lijevo) i dekoderom (desno).

Pretpostavimo kako na ulaz kodera dovodimo ulaznu rečenicu predstavljenu kao slijed vektorskih reprezentacija $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, a na njegovu izlazu dobivamo kodiranu reprezentaciju rečenice $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_n)$. Nakon toga, dekoder generira izlazni slijed $(y_1, y_2, \cdots, y_m)$, generirajući pritom izlaze slijedno (engl. *auto-regressive*). Ko-



Slika 3: Prikaz *Transformer* arhitekture. Slika nastala po uzoru na [13].

der modela *Transformer* sastoji se od $N = 6$ slojeva, pri čemu se svaki sloj kodera sastoji od dva podsloja. Prvi podsloj svakog sloja predstavlja više glava mehanizma pozornosti (engl. *multi-head self-attention*), dok je drugi podsloj jednostavna unaprijedna potpuno povezana neuronska mreža. Rezidualnim vezama (engl. *residual connections*) [14] ulazi svakog podsloja povezani su s njegovim izlazima, pri čemu se izlaz svakog podsloja zbraja s ulazom u taj podsloj te normalizira (engl. *layer normalization*) [15]. Zbog postojanja rezidualnih veza izlazi svakog podsloja imaju istu dimenziju $d_{model} = 512$. Kao i koder, dekoder modela *Transformer* sastoji se od $N = 6$ jednakih slojeva. Osim dva već spomenuta podsloja kodera, dekoder koristi i treći podsloj koji predstavlja više glava mehanizma pozornosti, a ulazi u taj sloj zapravo su kodirane reprezentacije dobivene na izlazu kodera (engl. *cross-attention*). Za svaki podsloj dekodera također se vežu rezidualne veze te normalizacija. U dekoderu je modificiran prvi podsloj mehanizma pozornosti tako da prilikom generiranja sljedeće riječi sloj pozornosti ne vidi buduće riječi, odnosno kažemo kako su buduće riječi maskirane (engl. *masked self-attention*). Ovakva modifikacija osigurava da dekoder u svakom koraku sljedeću riječ generira samo poznavajući sve riječi koje su se pojavile prije,

odnosno da dekoder doista modelira uvjetnu vjerojatnost:

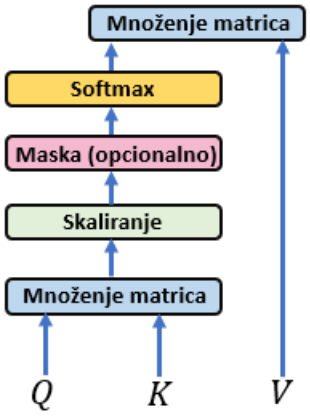
$$p\left(y_t \mid \{y_1, y_2, \cdots, y_{t-1}\}, \mathbf{Z}\right).$$

Ovaj mehanizam važan je za vrijeme učenje modela jer su nam za vrijeme iskorištavanja modela (engl. *inference*) i dostupne jedino dosad generirane riječi.

Funkcija koju obavljaju podslojevi mehanizma pozornosti nešto se razlikuje od funkcije opisane u odjeljku 2.3. Funkciju pozornosti kod *Transformer* arhitekture možemo opisati kao preslikavanje upita i skupa parova ključ-vrijednost u izlaz, pri čemu su upit, ključevi i vrijednosti vektori. Izlaz računamo kao težinsku sumu vektora vrijednosti, pri čemu težine računamo nekom funkcijom koja predstavlja sličnost između vektora upita i odgovarajućeg vektora ključa [13]. Zbog paralelizacije, u praksi se isplati računati funkciju pozornosti odjednom nad svim upitima. Zbog toga smještamo svaki vektor upita $\mathbf{q}$ (dimenzije d_q) u matricu upita $\mathbf{Q}$, a slično radimo i s vektorima ključeva $\mathbf{k}$ (dimenzije d_k) te vektorima vrijednosti $\mathbf{v}$ (dimenzije d_v). Mehanizam pozornosti temeljen na skaliranom skalarnom umnošku (engl. *scaled dot-product attention*) [13] koji koristimo u *Transformer* arhitekturi tada možemo predstaviti kao funkciju:

$$\textit{attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \textit{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{5}$$

U slučaju maskiranog mehanizma pozornosti dekodera, prije nego što provedemo funkciju *softmax*, na pozicije na koje ne želimo obraćati pozornost (pozicije veće od trenutne) postavljamo vrijednost $-\infty$ i tako efektivno sprječavamo model da vidi buduće riječi jer će pripadne težine nakon funkcije *softmax* težiti k 0. Mehanizam pozornosti temeljen na skalarnom umnošku ilustriran je na Slici 4.



Slika 4: Mehanizam pozornosti temeljen na skaliranom skalarnom umnošku [13].

Autori članka [13] opisali su i mehanizam pozornosti temeljen na više glava pozornosti, a takav podsloj možemo vidjeti na Slici 3. Najveća je razlika u tomu što su ulazi podijeljeni na više dijelova, pri čemu je svaka glava zadužena za pojedini dio. Osim toga, uvode se novi parametri u obliku matrica težina za svaku glavu: $\mathbf{W}_i^{\mathbf{Q}}, \mathbf{W}_i^{\mathbf{K}}, \mathbf{W}_i^{\mathbf{V}}$ te matrica težina za projekciju izlaza $\mathbf{W}^O$. Funkcija koju obavlja mehanizam pozornosti s h glava glasi:

$$\textit{multi-head}\,(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \textit{concat}\,(\textit{head}_1, \cdots, \textit{head}_h)\,\mathbf{W}^O,$$

pri čemu za funkciju $\textit{head}_i$ vrijedi:

$$\textit{head}_i = \textit{attention}\left(\mathbf{Q}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{K}\mathbf{W}_i^{\mathbf{K}}, \mathbf{V}\mathbf{W}_i^{\mathbf{V}}\right).$$

Budući da *Transformer*, za razliku od povratnih neuronskih mreža, nema povratne veze, gubi se informacija o redoslijedu riječi u rečenici. Zbog toga se u *Transformer* arhitekturi uvodi pozicijsko kodiranje (engl. *positional encoding*), pomoću kojega se pamti informacija o redoslijedu riječi unutar rečenice. Pozicijsko kodiranje temelji se na periodičnim funkcijama *sinus* i *kosinus*, koje se dodaju na svaku od d_{model} dimenzija vektorskih reprezentacija riječi iz ulazne rečenice.

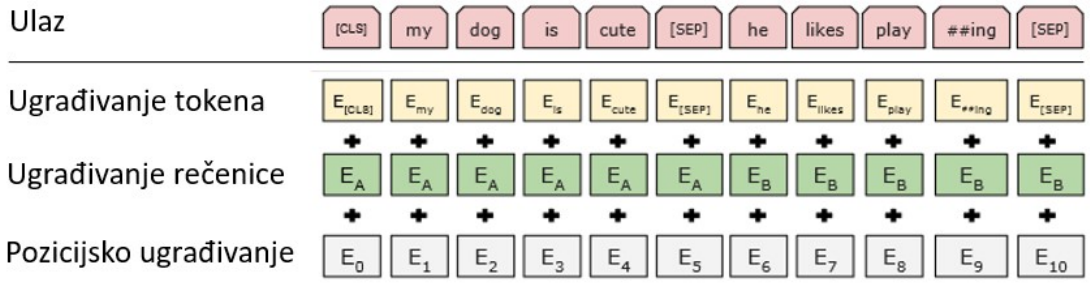
2.3.2 *BERT*

Model BERT (engl. *Bidirectional Encoder Representations from Transformers*) [16] višeslojni je dvosmjerni model koji se temelji na ranije opisanoj *Transformer* arhitekturi. Autori ovog rada koriste dva kriterija uspješnosti (engl. *objective function*) za vrijeme učenja modela. Prvi kriterij koji optimiziramo za vrijeme učenja previđanje je maskirane riječi (engl. *masked language modelling*, MLM). Model na ulazu dobiva rečenice iz skupa podataka za učenje, pri čemu se dio riječi maskira tokenom [MASK], a zadatak modela je da točno rekonstruira riječi koje se u skupu podataka nalaze na maskiranim mjestima. Autori rada slučajno su uzorkovali 15% riječi ulaznog teksta i 80% puta te riječi zamijenili [MASK] tokenom, 10% puta su riječ zamijenili drugom slučajno uzorkovanom riječi, dok 10% puta nisu radili nikakve preinake odabrane riječi. Osim predviđanja maskirane riječi, autori koriste i kriterij predviđanja sljedeće rečenice (engl. *next sentence prediction*, NSP), gdje je cilj za dvije ulazne rečenice predvidjeti nalaze li se jedna iza druge u ulaznom korpusu.

Ideja autora ovoga rada je prednaučiti model na velikoj količini podataka pomoću spomenutih kriterija uspješnosti, a zatim tako prednaučen model primijeniti na različite zadatke iz područja obrade prirodnog jezika. Cilj je da model u fazi predučenja uči kontekst u kojemu se pojedine riječi pojavljuju te da globalno dobro "razumije" jezik,

a nakon toga se tako prednaučeni model dodatno uči na različitim zadacima i manjim skupovima podataka. Dva su velika korpusa korištena u fazi predučenja: BooksCorpus [17] te korpus nastao preuzimanjem teksta s mrežne stranice *Wikipedia*.[2]

Pri učenju modela, dijelovi riječi ulazne rečenice preslikavaju se u slijed ugrađivanja *Wordpiece* (engl. *Wordpiece embeddings*) [18], a nakon toga se na ta ugrađivanja dodaju i pozicijska ugrađivanja (engl. *positional embeddings*) te ugrađivanja na razini rečenice (engl. *segment embeddings*). Slika 5 na primjeru prikazuje transformaciju ulaza pomoću spomenutih ugrađivanja. U okviru ovog rada model BERT-a koristimo za učenje klasifikatora crta osobnosti. Korišteni model zapravo je temeljen na koderu *Transformer* arhitekture, a ima 12 slojeva, veličina je reprezentacija 768, broj glava u slojevima pozornosti jest 12, a model ukupno ima oko 110 milijuna parametara.



Slika 5: Primjer transformacije ulaza pomoću ugrađivanja.

2.3.3 *GPT2*

Model GPT2 [19] jednosmjerni je višeslojni model koji se temelji na dekoderu *Transformer* arhitekture. Cilj učenja ovog modela predviđanje je sljedeće riječi (engl. *language modelling*) na temelju riječi iz prošlosti. Arhitektura modela slična je dekoderu modela *Transformer*, uz uklanjanje podsloja pozornosti uvjetovanog izlazima kodera te postavljanje normalizacije na ulaz u svaki podsloj. Kao i kod modela *Transformer*, dekoder smije vidjeti samo riječi iz prethodnih vremenskih trenutaka, zbog čega se kao prvi podsloj svakog sloja koristi mehanizam maskirane pozornosti. Ideja je na zadatku predviđanja sljedeće riječi prednaučiti model na velikoj količini podataka. Nakon što model nauči kontekst u kojemu se riječi pojavljuju, ideja je taj model primijeniti na različite zadatke u okviru područja obrade prirodnog jezika. U svrhu predučenja modela, autori rada stvorili su skup podataka WebText, a nakon filtriranja skup podataka

[2] `https://www.wikipedia.org/`

je imao oko osam milijuna dokumenata i 40GB teksta. Zbog postizanja kompromisa između predviđanja sljedećeg znaka i predviđanja sljedeće riječi, autori rada za kodiranje ulaza koriste kodiranje digrama (engl. *byte pair encoding*, BPE) [20]. U okviru ovoga rada model GPT2 koristimo kao osnovni prednaučeni model za generiranje razgovora.

2.4 Konverzacijski modeli

Konverzacijski modeli omogućuju ljudima interakciju s računalom na jednak način kao što komuniciraju i s drugim ljudima. Najčešće se pojavljuju kao konverzacijski modeli koji odgovaraju na korisnikova postavljena pitanja te mu pomažu u rješavanju problema. Postoje i konverzacijski modeli koji mogu voditi svakodnevni razgovor s osobom o širokom spektru tema te s neograničenim brojem odgovora. Razvoj konverzacijskih modela u počecima se temeljio na korištenju arhitektura RNN, LSTM i GRU, a u novije vrijeme temelji se na *Transformer* arhitekturi.

Autori rada [21] kodirali su dvosmjernim modelom LSTM izjave razgovora uz oznake domene i sentimenta. Njihov model temelji se na arhitekturi LSTM te može voditi sentimentom obogaćen razgovor s osobom. U članku [22] autori su koristili arhitekturu GRU za stvaranje atributa svojstvenih korisniku. Njihov konverzacijski model može učinkovito učiti informacije o korisniku kroz vrijeme. Autori rada [22] svoj model temelje na arhitekturi GAN [23], koja se primarno koristi pri generiranju slika. Kako bi učinkovito pamtili prošlost razgovora, autori rada [24] dizajnirali su model THRED, koji se temelji na arhitekturi GRU za kodiranje prošlosti i modelu LDA [25] za ekstrakciju spomenutih tema u razgovoru. Pojavom *Transformer* arhitekture, autori rada [26] koristili su prethodno naučen model BERT za kodiranje prošlosti razgovora. U članku [27] autori koriste model BERT kao podlogu za konverzacijski model i kodiranje prošlosti razgovora. Nakon dodatnog učenja modela na tekstovima preuzetih s *Reddita* i *Twittera*, autori su naučili te ispitali model na više konverzacijskih skupova podataka te uspjeli postići dotad najbolje rezultate. Pored modela BERT, pojedini radovi koriste model GPT2 za stvaranje konverzacijskih modela [28]. Takav je i model DialoGPT [29] nastao dodatnim učenjem modela GPT2 nad razgovorima s *Reddita*.

2.4.1 *DialoGPT*

Model DialoGPT temelji se na arhitekturi modela GPT2. Dodatno, autori su učili GPT2 na razgovorima preuzetim s foruma *Reddit*, koji su se odvili u vremenu između 2005. i 2017. godine. Kako bi formirali razgovore, iskoristili su komentare objava

koji međusobno odgovaraju jedni na druge. Filtrirali su one razgovore koji sadrže poveznice, uzastopno ponavljanje tri riječi, ne sadrže zaustavnu riječ engleskog jezika, posjeduju jezik za označavanje (engl. *markup language*), duži su od 200 riječi ili sadrže neprimjerene riječi. Dodatno, uklonili su razgovore koji sadrže 90% najčešćih 3-grama u skupu dijaloga. Konačni skup dijaloga imao je oko 147 milijuna razgovora s preko 1.8 milijardi riječi. Naizmjenične izjave dijaloga spojili su u jedinstveni tekst $(\mathbf{x}_1, \cdots, \mathbf{x}_N)$ te unosili u model GPT2. Cilj modela maksimizirati je vjerojatnost:

$$p(T \mid S) = \prod_{n=m+1}^{N} p\left(\mathbf{x}_n \mid \mathbf{x}_1, \cdots, \mathbf{x}_{n-1}\right),$$

pri čemu su prethodne izjave prikazane kao $S = (\mathbf{x}_1, \cdots, \mathbf{x}_m)$, dok je trenutni odgovor prikazan kao $T = (\mathbf{x}_{m+1}, \cdots, \mathbf{x}_N)$. Naučeni model postigao je dotad najbolje rezultate na skupu podataka DSTC-7 [30], a generirani dijalozi bili su vrlo slični ljudskima.

2.5 Usmjeravanje modela korištenjem predložaka

U nadziranom učenju nastoji se povećati vjerojatnost točne klase $p(y|\mathbf{x};\boldsymbol{\theta})$ na temelju ulaznog teksta $\mathbf{x}$ i parametara modela $\boldsymbol{\theta}$. Metoda usmjeravanja modela pomoću predložaka (engl. *prompt-based learning*) [31] pristupa učenju na drugačiji način. Ona nastoji povećati vjerojatnost samog teksta $p(\mathbf{x};\boldsymbol{\theta})$ izgradnjom predloška oko prvobitnog ulaznog teksta. Ovim načinom bolje se iskorištavaju modeli koji su prednaučeni upravo u svrhu povećavanja vjerojatnosti ulaznog teksta $p(\mathbf{x})$.

Tablica 1: Potpuni postupak stvaranja predloška i terminologija

Ime	**Notacija**	**Primjer**	**Opis**
Ulaz	**x**	Volim ovaj film	Jedan ili više tekstova
Izlaz	**y**	++ (vrlo pozitivan)	Klasa ili tekst
Funkcija za izradu predloška	$f_{predložak}(\mathbf{x})$	$[X]$ Sveukupno, bio je to $[Z]$ film.	Funkcija koja pretvara ulaz u specifičan oblik unošenjem ulaza **x** i stavljajući znak $[Z]$ gdje će se unijeti točan odgovor **z**.
Predložak	**x'**	Volim ovaj film. Sveukupno, bio je to $[Z]$ film.	Tekst gdje je $[X]$ odgovoren s ulaznim podatkom **x**, a znak za odgovor $[Z]$ nije.
Ispunjen Predložak	$f_{ispuni}(\mathbf{x'}, \mathbf{z}^*)$	Volim ovaj film. Sveukupno, bio je loš film.	Predložak gdje je znak $[Z]$ zamijenjen nekim odgovorom.
Odgovoren Predložak	$f_{ispuni}(\mathbf{x'}, \mathbf{z})$	Volim ovaj film. Sveukupno, bio je dobar film.	Predložak gdje je znak $[Z]$ zamijenjen točnim odgovorom.
Odgovor	**z**	"dobar", "odličan", "loš"	Riječ, izraz ili rečenica koja mijenja $[Z]$.

Kod izrade predloška za učenje modela, koristi se funkcija za izradu predloška $f_{predložak}$. Predložak se izrađuje u dva koraka. U prvom koraku dodaju se riječi predloška i umeću znakovi [X] i [Z] na odgovarajuća mjesta, ovisno kakav predložak stvaramo. U drugom koraku na mjesto [X] postavlja se tekst iz podataka, a na mjesto [Z] traži se od modela da unese točan odgovor koji se preslikava u točnu klasu. Primjerice, za klasifikaciju sentimenta možemo koristiti odgovore {"odlično","dobro","OK", "loše","užasno"} koji se preslikavaju u točne klase {++, +, / , -, –}. Potpuni postupak i terminologija stvaranja predložaka prikazana je u Tablici 1.

Stvaranju predloška pristupa se na više načina. Jedan pristup je stvoriti predložak ručno na temelju uvida u zadatak. Kako je ručno stvaranje predloška vrlo teško i često ne rezultira optimalnim predlošcima, razvile su se automatske metode stvaranja predložaka. Automatske metode stvaraju predloške koji se dijele na diskretne i kontinuirane, ovisno o tome koristi li se tekst ili gusta vektorska reprezentacija.

Tekstni predložak je nastao automatskom metodom pretrage za optimalnim predloškom. Jedna metoda stvaranja diskretnog predloška akvizicija je predložaka, koja pomoću ulaznog teksta **x** i stvarne klase **y** ili odgovora **z** nastoji pronaći riječi ili skupine riječi u korpusu teksta koje će postati dio predloška. Metoda temeljena na parafraziranju uzima postojeći predložak i parafraziranjem pronalazi slične predloške. Pojedine metode koriste jezične modele koji generiraju potpuno nove predloške ili dovršavaju započete predloške. Stvoreni predlošci mogu se ocijeniti pomoću jezičnog modela kako bi se saznalo koji je predložak najpogodniji. Modelu se unosi predložak i na temelju izlaznih vjerojatnosti izračunava se ukupna vjerojatnost predloška. Predložak s najvećom vjerojatnošću potencijalno je optimalan predložak.

Kontinuirani predložak ne predstavlja tekst. Kontinuirani predložak skup je vektora koji se dodaju na određena mjesta oko ulaznog teksta. Dodani vektori se uče, dok se vektori riječi i parametri cjelokupnog modela drže fiksnima. Učeći isključivo dodane vektore, nastoji se izvući znanje modela naučeno tijekom faze predučenja. Postoji nekoliko pristupa učenju kontinuiranog predloška. Prefiksno učenje najkorišteniji je pristup koji dodaje vektore predloška na početak vektora reprezentacije riječi u rečenici. Kako bi se dodani vektori što bolje inicijalizirali, inicijalizacija se obično vrši pomoću vektora riječi već poznatog tekstnog predloška, vektorima točnih klasa, slučajnim uzorkovanjem vektora iz vokabulara vektora modela ili drugim metodama.

Osim stvaranja samog predloška, potrebno je stvoriti i odgovore predloška koji će biti preslikani u točne klase. Kod stvaranja odgovora predloška, potrebno je definirati oblik odgovora i njegov dizajn. S obzirom na oblik koji poprima, odgovor predloška može biti riječ, skup riječi ili rečenica. Riječi ili skupovi riječi obično se koriste u klasifikaciji, a rečenice se koriste u zadacima vezanim uz generiranje teksta.

Pri stvaranju dizajna odgovora predloška postoje različite metode. Metoda ručnog stvaranja predloška stvara odgovor na temelju uvida u zadatak i stvoreni predložak te na temelju domenskog znanja. Kod stvaranja odgovora, obično se definira je li skup mogućnosti ograničen ili neograničen. Ograničen skup odgovora koristi se u klasifikacijskim zadacima gdje postoji samo nekoliko klasa u koje preslikavamo odgovor. U zadacima generiranja, skup je mogućih odgovora beskonačan. Ručno je stvaranje odgovora teško, stoga se koriste metode automatskog pronalaska prikladnog odgovora predloška. Parafraziranje odgovora predloška koristi parafraziranje kako bi se proširio broj mogućnosti odgovora na predložak. Dobiveni skup odgovora ispituje se na temelju performansi modela kako bi se izabrao skup najboljih odgovora za korištenje. Postoje metode gdje se prikladan odgovor pronalazi na temelju učenja vektora odgovora, ali takve metode rijetko se koriste u praksi.

Ovisno o tomu učimo li model, dodaje li predložak nove parametre i uči li se predložak, postoje različiti pristupi učenju modela. Učenje bez predložaka uči modele u klasičnoj okolini gdje se modelu predaje tekst, a od njega traži da na izlazu postavi vjerojatnosti pripadnosti klasama zadatka. Korištenje fiksiranog modela s predloškom metoda je koja koristi najčešće ručno stvorene predloške kako bi iskoristila znanje prednaučenog modela u rješavanju zadatka. Model ispunjava predložak odgovorom koji se preslikava u klase zadatka. Učenje predloška uz fiksirani model metoda je koja koristi kontinuirane predloške za rješavanje zadatka. Kontinuirani predlošci unose nove parametre, ali broj parametara koji se uči mnogo je manji jer se model drži fiksiranim. Učenje modela pomoću predloška koristi predložak za učenje modela gdje se povećava vjerojatnost modela da ispuni predložak odgovarajućim odgovorom. Učenje modela i predloška koristi kontinuirane predloške, ali uči se i sami model.

Područje učenja modela pomoću predložaka široko je te se broj radova u području eksponencijalno povećava. Autori članka [32] stvorili su više predložaka ispunjavanjem maskiranih mjesta u tekstu riječima pomoću modela T5 [33]. U članku [34] autori uvode korak posredničkog učenja gdje su prikupili skupove podataka od više zadataka te ih izmijenili da budu u obliku predloška s da/ne odgovorom. Nakon učenja

modela nad skupom sličnih zadataka, autori su testirali model bez učenja na zadatku koji pripada tom skupu te dobili bolje performanse od nenaučenog modela. U radu [35] autori su pokazali da potraga za hiperparametrima i optimalnim predlošcima u učenju s nekoliko primjera nije bolja od slučajnog odabira. Autori članka [36] generirali su predloške pomoću modela T5. Pomoću generiranih predložaka i dodavanja nekoliko demonstracija ispunjenih predložaka prije neispunjenog predloška, postigli su bolje performanse nego standardno učenje na nekoliko primjera. Autori rada [37] dodatno su prednaučili model T5 100 tisuća koraka kako bi ga koristili pomoću predložaka. Mijenjali su duljinu predloška, inicijalizaciju predloška, učenje cijelog modela ili samo predloška. Učenje kontinuiranog predloška pokazalo se kao dobar pristup za korištenje velikih modela na različitim zadacima. Autori članka [38] pokazali su kako korištenje jednostavnih predložaka, gdje su tekst i odgovor samo konkatenirani, te kontinuiranih predložaka, daje jednake rezultate prilikom učenja na nekoliko primjera. Ručno izrađeni predlošci pokazali su se najboljima. Autori rada [39] koristili su različite veličine modela T5 s prefiksnim učenjem. Koristeći prefikse s 20 do 100 posebnih znakova preslikanih u kontinuirane predloške, postigli su bolje rezultate od modela GPT3 i T5 XXL učenih na samo nekoliko primjera. U radu [40] autori su koristili prefiksno učenje. Učeći isključivo kontinuirani predložak, s 0.1% učenih parametara postigli su jednake ili bolje rezultate na tri zadatka vezana uz generiranje teksta.

3 Osobnost

Konverzacijski modeli koje ćemo stvoriti, posjedovat će sposobnost vođenja razgovora s čovjekom i jedinstvenu crtu osobnosti. Univerzalna definicija osobnosti ne postoji, ali većina teorija fokusira se na motivaciju i psihološke karakteristike inducirane od strane okoline. Pojedine teorije opisuju osobnost kao više crta osobnosti na temelju kojih se može predvidjeti ponašanje. Druge metode koje se temelje na ponašanju definiraju osobnost kroz učenje izvana i navike. Najpoznatiji i najčešće korišteni modeli osobnosti su Big Five i MBTI.

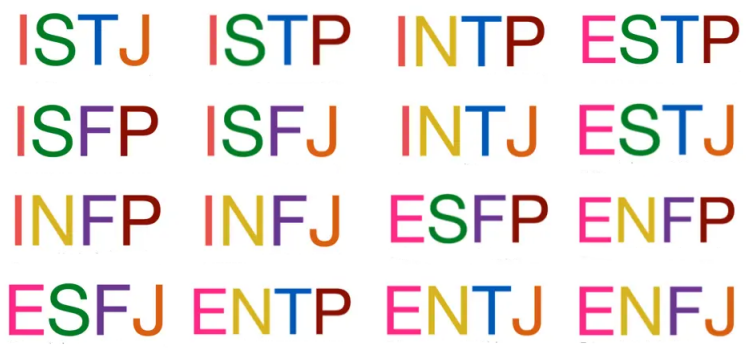
3.1 Model osobnosti Big Five

Model osobnosti Big Five predstavlja osobnost kroz pet crta. Prva je crta osobnosti otvorenost (engl. *openness*), koju povezujemo sa zanimanjem za umjetnost, avanturu, emocije, maštu i nova iskustva. Osobe s izraženom crtom otvorenosti obično su kreativnije i svjesne svojih osjećaja. Takve osobe avanturistički su nastrojene te sklonije nekonvencionalnim razmišljanjima. S druge strane, osobe niske otvorenosti ustrajnije su, tradicionalnog karaktera te više cijene činjenice. Druga je crta osobnosti savjesnost (engl. *conscientiousness*), koju povezujemo uz samodisciplinu i ustrajnost u ostvarenju vlastitih ciljeva neovisno o vanjskim preprekama. Osobe visoke savjesnosti doimaju se tvrdoglavima i fokusiranima, dok su osobe niske savjesnosti fleksibilne i spontane. Treća je crta osobnosti ekstravertiranost (engl. *extraversion*), koju povezujemo uz sklonost izlaganju društvenim događanjima te gubljenju ili crpljenju energije iz njih. Ekstravertirane osobe vole provoditi vrijeme u društvu i doimaju se punima energije. Introvertirane osobe, s druge strane, povučenije su te preferiraju provoditi vrijeme u samoći. Četvrta je crta osobnosti prijaznost (engl. *agreeableness*), koju povezujemo s težnjom osobe za društvenom skladnošću. Osobe s izraženom crtom prijaznosti dobro se slažu s drugima, drage su, može im se vjerovati i spremne su se prilagoditi. Niska prijaznost odlikuje se u fokusu na osobne potrebe, a takve osobe natjecateljski su nastrojene te se ponekad mogu doimati kao svadljive ili kao osobe kojima se ne može vjerovati. Posljednja je crta osobnosti neuroticizam (engl. *neuroticism*), a povezujemo ju sa sklonosti proživljavanju negativnih emocija poput ljutnje, anksioznosti ili depresije. Na Slici 6 možemo vidjeti pet crta osobnosti prema modelu Big Five.

Slika 6: Crte osobnosti prema modelu osobnosti Big Five [41].

3.2 Model osobnosti MBTI

Model osobnosti MBTI introspektivan je model za samoprocjenu osobnosti koji opisuje čovjekovu osobnost prema četiri oprečne crte osobnosti: ekstravertiranost (engl. *extraverted*) i introvertiranost (engl. *introverted*), raspoznavanje (engl. sensing) i intuitivnost (engl. *intuition*), razmišljanje (engl. *thinking*) i osjećanje (engl. *feeling*) te opažanje (engl. *perceiving*) i prosuđivanje (engl. *judging*). Ekstrovertirani pojedinci skloni su akciji i sudjelovanju, dok introvertirani pojedinci troše vlastitu energiju kroz akciju i sudjelovanje te trebaju vrijeme za sebe kako bi obnovili izgubljenu energiju. Raspoznavanje i intuicija povezani su s načinom reagiranja na informacije. Osobe s izraženijom crtom osobnosti raspoznavanja preferiraju informacije u sadašnjosti, opipljive i konkretne. Intuitivni pojedinci manje su koncentrirani na svoja trenutna osjetila i okrenuti su prema budućnosti. Ovisno o načinu donošenja odluka, osobe su sklonije crti osobnosti razmišljanja ili osjećanja. Pojedinci skloni razmišljanju donose odluke na temelju logike, prateći pravila i uzimajući u obzir uzročno posljedične veze. Osjećajniji pojedinci donose odluku na temelju situacije i uključenih pojedinaca kako bi se postiglo najbolje rješenje. Ovisno o tom kako osoba doživljava svijet, ona poprima crtu osobnosti opažanje ili prosuđivanje. Osobe koje se odnose prema vanjskom svijetu primarno s crtom osobnosti raspoznavanja ili intuicije su opažajne osobe. Osobe s izraženim crtama razmišljanja ili osjećanja osobe su sklone prosuđivanju. Ovisno o pripadnosti pojedinim crtama, postoji 16 osobnosti koje se najčešće prikazuju pomoću početnih slova svojih crta osobnosti te su prikazane na Slici 7.



Slika 7: Crte osobnosti prema modelu osobnosti MBTI [42]

3.3 Predviđanje osobnosti

Među prvim istraživanjima u predviđanju osobnosti bio je rad [43]. Autori tog rada pokušali su povezati korištenje jezika i osobnosti. Nekoliko radova koncentriralo se na automatsko predviđanje osobnosti iz e-pošte [44], eseja [45], govora izvučenih iz snimki [46], objava na *Twitteru*[3] [47] te različitih blogova [48]. Skup podataka MyPersonality [49] bio je prvi skup podataka s označenim osobnostima dobiven s društvenih mreža. Imao je označene osobnosti po modelu Big Five za 7.5 milijuna korisnika društvene mreže *Facebook*[4]. Autori rada [50] iskoristili su spomenuti skup podataka kod stvaranja modela za predviđanje osobnosti na temelju teksta. Tijekom stvaranja modela prikupili su 15.4 milijuna statusa pojedinaca. U radovima [51, 52] stvorena su dva zadatka temeljena na skupovima podataka označenim osobnostima po modelu Big Five. Istraživanja na ovim zadacima pokazala su kako postupci koji se temelje na n-gramima riječi daju najbolje rezultate. U članku [53] autori su prikupili 1.2 milijuna statusa s *Twittera* s označenim osobnostima MBTI. Skup podataka PANDORA [2] jedan je od najnovijih skupova podataka prikupljen s foruma *Reddit*. Skup podataka nastao je kao nadogradnja na rad [54] i sadrži označene osobnosti modela MBTI i Big Five za više od 10 tisuća pojedinaca te 17.5 milijuna komentara. Pojavom dubokog učenja, većina članaka o predviđanju osobnosti okrenula se dubokim pristupima [55, 56, 57].

[3] `https://twitter.com/`
[4] `https://hr-hr.facebook.com/`

3.4 Skup podataka PANDORA

Skup podataka PANDORA [2] prvi je skup podataka s *Reddita* označen osobnostima MBTI i demografskim podacima. Autori skupa podataka prikupili su 17.5 milijuna komentara od preko 10 tisuća korisnika. Za svakog korisnika dodatno su prikupili osobnosti po modelu Big Five, Enneagram[5] i MBTI te demografske podatke poput starosti, spola, lokacije i jezika. Skup podataka stvoren je kao nadogradnja na rad [54], u kojemu su autori prikupili MBTI osobnosti korisnika na temelju oznaka uz korisničko ime (engl. *flair*) u podforumima na *Redditu* vezanim uz model osobnosti MBTI. Prikupljena je osobnost za 9054 korisnika, a autori skupa podataka PANDORA na sličan su način dodatno prikupili 793 Enneagram osobnosti. Tablica 2 prikazuje MBTI crte osobnosti korisnika.

Tablica 2: Distribucija crta osobnosti MBTI u skupu podataka PANDORA

MBTI Dimenzija	Korisnici	MBTI Dimenzija	Korisnici
Introvertiranost	7134	Ekstravertiranost	1920
Intuitivnost	8024	Raspoznavanje	1030
Razmišljanje	5837	Osjećanje	3217
Opažanje	5302	Prosuđivanje	3752

Prikupljanje osobnosti po modelu Big Five bilo je teže jer su one objavljene u komentarima korisnika. Dodatno, osobnosti po modelu Big Five prikazane su u obliku brojčanih vrijednosti za pojedine crte osobnosti te postoje različiti upitnici za njihovo prepoznavanje. Nakon prikupljanja komentara koji su sadržavali dobro napisane crte osobnosti, uvidom u objavu od pojedinih komentara detektirani su upitnici osobnosti na koje se komentar odnosio. Na ovaj način uspješno je pronađeno 1027 osobnosti po modelu Big Five. Kako pojedine objave nisu jasno prikazale na koji upitnik se odnosi, izgrađen je n-gramski model na temelju poznatih objava za predviđanje upitnika osobnosti nepoznatih objava. Pomoću ovog modela autori su dodatno pronašli još 600 osobnosti korisnika. Srednje vrijednosti osobnosti po modelu Big Five prikazane su u Tablici 3, pri čemu je za pojedinu crtu osobnosti minimalna vrijednost 0, a maksimalna vrijednost 100, što odgovara potpunom nepripadanju odnosno pripadanju crti osobnosti.

[5] `https://en.wikipedia.org/wiki/Enneagram_of_Personality`

Tablica 3: Srednje vrijednosti crta osobnosti korisnika po modelu Big Five

Big Five Dimenzija	Svi	Žene	Muškarci
Otvorenost	62.5	62.9	64.3
Savjesnost	40.2	43.3	41.6
Ekstravertiranost	37.4	39.7	37.6
Prijaznost	42.4	44.1	38.9
Neuroticizam	49.8	51.6	46.9

4 Eksperimenti

4.1 Priprema skupova podataka

Početni skup podataka PANDORA sadrži 17.5 milijuna komentara. Za učenje klasifikatora crta MBTI osobnosti koji će biti korišteni kasnije u filtraciji dijaloga s *Reddita*, početni je broj komentara prevelik. Karakteristike osobnosti nisu očite u većini komentara u skupu podataka, što može utjecati na performanse naučenih klasifikatora. Dodatno, broj komentara prevelik je kako bi se u efektivnom vremenu naučile osnovne veličine modela. Kako bismo naučili modele sa što boljim performansama, stvorili smo manje skupove podataka za svaki par crta osobnosti.

U prvom koraku filtrirali smo sve komentare čiji autor ne posjeduje cjelokupnu MBTI osobnost u PANDORA skupu podataka. Nakon filtriranja ostalo je 15.5 milijuna komentara od 9067 korisnika. Budući da ćemo učiti klasifikatore da predviđaju crte osobnosti iz samog komentara, uparili smo svaki komentar s odgovarajućim crtama osobnosti osobe. Uvidom u komentare uočili smo kako velik broj komentara sadrži tekstove koji ne podsjećaju na prirodan govor engleskog jezika, sadrže mnoštvo poveznica i automatske tekstove moderatora. Kako bismo uklonili komentare sa spomenutim karakteristikama, filtrirali smo sve komentare koji sadrže uzorke iz Tablice 4 i ne sadrže nijednu zaustavnu riječ engleskog jezika. Nakon dodatnog filtriranja komentara koji imaju manje od 5 ili više od 256 riječi, preostalo nam je 12.1 milijuna komentara. Distribucija crta osobnosti nad komentarima nakon filtriranja prikazana je u Tablici 5.

Tablica 4: Primjeri korištenih uzoraka za filtriranje komentara

Uzorci
https://
http://
Thank you for your submission!
This comment or submission has been removed
All replies to this post must be a maximum
Here you can write whatever!
^^^^^^^

Tablica 5: Distribucija crta MBTI osobnosti nad komentarima

MBTI Dimenzija	Komentari	%	MBTI Dimenzija	Komentari	%
Introvertiranost	9.6M	79.3	Ekstravertiranost	2.5M	20.7
Intuitivnost	10.5M	86.8	Raspoznavanje	1.6M	13.2
Razmišljanje	8.9M	73.6	Osjećanje	3.2M	26.4
Opažanje	7M	57.8	Prosuđivanje	5.1M	42.2

Početni skup komentara za svaki par crta osobnosti podijelili smo u skup za učenje s dvije trećine te u ispitni skup s jednom trećinom komentara. Nakon što smo dobili prednaučene modele na skupovima za učenje, testirali smo ih na ispitnim skupovima. Kao evaluacijsku metriku koristimo makro F_1 vrijednost. Makro F_1 vrijednost računa se kao harmonijska sredina preciznosti i odziva prema izrazu:

$$F_1 = \frac{2 \times \text{ preciznost } \times \text{ odziv}}{\text{preciznost } + \text{ odziv}}.$$

Preciznost računamo kao omjer točno klasificiranih pozitivnih primjera i ukupnog broja pozitivno klasificiranih primjera, dok odziv računamo kao omjer točno klasificiranih pozitivnih primjera te ukupnog broja pozitivnih primjera. Kako bismo ravnopravno gledali sve klase kao pozitivne, koristimo makro F_1 vrijednost, koja usrednjuje F_1 vrijednost za svaku od klasa.

Tijekom testiranja, za svaki komentar iz ispitnog skupa izvukli smo pouzdanosti modela u obliku vjerojatnosti za klasificiranje pozitivne crte osobnosti iz para. Za pozitivne osobnosti izabrali smo introvertiranost, intuitivnost, razmišljanje i opažanje. Eksperimente smo ponovili tri puta uzimajući uvijek nove komentare u trećinski ispitni skup. Svaki komentar na kraju je eksperimenata imao četiri vjerojatnosti za pozitivne crte osobnosti iz četiri para oprečnih crta osobnosti.

Za učenje koristimo model TinyBERT [58] s 14.5 milijuna parametara te četiri sloja *Transformer* arhitekture, 12 glava pozornosti i veličinom vektorske reprezentacije 312. Skup za učenje dodatno smo podijelili na skup za učenje s 80% komentara te na skup za validaciju s 20% komentara. Modele smo učili 10 epoha koristeći pritom rano zaustavljanje na temelju performansi modela na validacijskom skupu i sa spremanjem najboljeg modela tijekom učenja. Koristili smo 16 primjera za veličinu grupe i optimizator AdamW [59]. Za stopu učenja izabrali smo vrijednost $2e$-5 na temelju validacijskih performansi modela tijekom učenja na paru introvertiranost i ekstravertiranost. Koristili smo linearno smanjivanje stope učenja te smo komentare prebacili u

mala slova. Za implementaciju koristimo `Pytorch` [60], a sav računalni kod eksperimenata dostupan je na *Githubu*.[6]

Prvi rezultati bili su poprilično loši zbog izrazite neuravnoteženosti skupa podataka. Kako bismo uravnotežili početni skup podataka, slučajno smo uzorkovali 1.5 milijuna komentara za svaku opručnu crtu osobnosti, što je rezultiralo s četiri skupa podataka od kojih se svaki sastojao od 3 milijuna komentara. Usrednjene makro F_1 vrijednosti za tri pokretanja i svaki par crta osobnosti prikazani su u Tablici 6.

Tablica 6: Rezultati učenja modela TinyBERT

Crte osobnosti	F_1
Ekstravertiranost/Introvertiranost	58.18
Raspoznavanje/Intuitivnost	60.08
Razmišljanje/Osjećanje	59.92
Prosuđivanje/Opažanje	56.25

x Koristeći pouzdanost modela za pozitivne crte osobnosti, stvorili smo četiri manja skupa podataka za učenje osnovnih veličina modela za klasifikaciju osobnosti. Komentari koji sadrže znatno veću ili manju pouzdanost od 50% vjerojatno sadrže očitije značajke potrebne za predviđanje crta osobnosti, stoga smo ostavili one komentare koji imaju veću pouzdanost od pouzdanosti prikazanih u Tablici 7. Pouzdanost za negativne crte osobnosti prikazali smo kao 100% – pouzdanost pozitivne crte osobnosti. Nakon filtriranja, dobili smo četiri skupa podataka, od kojih svaki sadrži 300 tisuća komentara s distribucijom osobnosti prikazanom u Tablici 7.

Tablica 7: Granice pouzdanosti modela za filtriranje komentara

Dimenzija	Broj %	Granica %	Dimenzija	Broj %	Granica %
Ekstravertiranost	53.12	74.9	Introvertiranost	46.88	72.6
Raspoznavanje	58.04	94.5	Intuitivnost	41.96	84.8
Osjećanje	48.86	76.3	Razmišljanje	51.14	76.4
Prosuđivanje	51.86	68.6	Opažanje	48.14	68.0

[6] `https://github.com/helen1c/chatbot/`

4.2 Učenje klasifikatora crta osobnosti

Za učenje klasifikatora crta osobnosti izabrali smo osnovnu veličinu modela BERT sa 110 milijuna parametara, 12 slojeva *Transformer* arhitekture, veličinom vektora stanja 768 te s 12 glava pozornosti. Stvorena četiri skupa podataka podijelili smo na skup za učenje sa 70% komentara, validacijski skup s 10% komentara te ispitni skup s 20% komentara. Učili smo model kroz 6 epoha s ranim zaustavljanjem na temelju validacijske makro F_1 vrijednosti. Koristili smo optimizator AdamW s linearnim smanjivanjem stope učenja. Veličina grupe podataka bila je 2, zbog ograničenih memorijskih resursa, a komentare smo pretvorili u mala slova prije unošenja u model. Kako bismo odabrali vrijednost stope učenja, učili smo model sa stopama učenja $5e$-5, $2e$-5, $1e$-5 i $5e$-6 na paru s crtama osobnosti introvertiranost i ekstravertiranost. Najbolje rezultate na validacijskom skupu dobili smo sa stopom učenja $1e$-5. Rezultati naučenih modela na ispitnim skupovima prikazani su u Tablici 8. Model za klasifikaciju crta osobnosti raspoznavanje i intuitivnost pokazao se kao najbolji s visokom makro F_1 vrijednosti od 84.76%. Najmanju makro F_1 vrijednost od 70.44% na ispitnom skupu ostvario je model za klasifikaciju crta osobnosti prosuđivanje i opažanje. Naučene modele koristimo kako bismo filtrirali dijaloge s *Reddita* u osam skupova dijaloga s različitim MBTI crtama osobnosti.

Tablica 8: Performanse klasifikatora crta osobnosti

Crte osobnosti	F_1
Ekstravertiranost/Introvertiranost	75.84
Raspoznavanje/Intuitivnost	84.76
Razmišljanje/Osjećanje	77.94
Prosuđivanje/Opažanje	70.44

4.3 Preuzimanje i obrada dijaloga s *Reddita*

Kako bismo naučili modele koji će posjedovati crtu osobnosti, potreban nam je skup dijaloga koji ćemo filtrirati te iskoristiti za učenje modela. Za stvaranje skupa dijaloga iskoristili smo forum *Reddit*. Stvorili smo dijaloge tako što smo preuzeli objave određenog dana i podforuma (engl. *subreddit*) te smo komentare na objavu, koji slijedno odgovaraju jedni na druge, pretvorili u dijalog. Za podforume iz kojih ćemo uzorkovati objave, odabrali smo listu od 4074 najpopularnija podforuma u vrijeme našeg pristupanja *Redditu*. Pojedini podforumi prikazani su u Tablici 9. U skupu podataka postojali

su podforumi koji posjeduju neprimjeren sadržaj, stoga smo takve podforume filtrirali pomoću liste 549 poznatih neprimjerenih podforuma.

Tablica 9: Primjeri podforuma za stvaranje dijaloga

Podforumi	
politics	memes
gaming	MadeMeSmile
leagueoflegends	funny
Minecraft	explainlikeimfive
nba	europe
news	anime
videos	Music

Uzorkovali smo objave za dane između 1. 1. 2018. i 10. 3. 2022. Raspon dana izabran je kako bi teme koje se obrađuju bile novije, a tekst za učenje se ne bi preklapao s tekstom pomoću kojeg je korišteni konverzacijski model bio prednaučen. Za preuzimanje objava i komentara, koristili smo biblioteku `PRAW` [61] i *Reddit* API [62]. Pomoću *Reddit* API-ja preuzeli smo identifikatore objava određenog dana i podforuma, a pomoću biblioteke `PRAW` efektivno smo povukli same objave i komentare. Biblioteka `PRAW` omotač je oko *Reddit* API-ja koja učinkovito rješava probleme vezane uz ograničen broj pristupa u određenom vremenu. Kako bismo što učinkovitije preuzeli i stvorili dijaloge, pokrenuli smo 350 potprocesa pri čemu je svaki potproces zadužen za stvaranje dijaloga za manji dio ukupne liste podforuma. Format dijaloga pratio je željeni format modela kojeg smo koristili za učenje. Nekoliko stvorenih dijaloga prikazano je u Tablici 10. Oznaka 0.0 označava ono što je korisnik rekao, a oznaka 1.0 označava ono što bi model trebao naučiti reći. Ukupno smo prikupili oko 81 milijun dijaloga s *Reddita*, koji su dostupni na poveznici.[7]

Pojedini dijalozi sadržavali su neprimjerene riječi koje ne želimo da model nauči koristiti. Na temelju liste od 75 najčešće korištenih neprimjerenih riječi, filtrirali smo sve dijaloge u kojima se bar jednom pojavila takva riječ. Veliki broj dijaloga sadržavao je poveznice i jezik za označavanje, što je modelu beskorisno, pa smo takve dijaloge uklonili. Postojali su i dijalozi s mnogo ponavljanja, kao i oni koji ne izgledaju kao prirodan govor osobe. Kako bismo filtrirali dijaloge ove vrste, uklonili smo sve dijaloge koji sadrže uzastopno ponavljanje od barem 3 riječi ili ne sadrže nijednu

[7] `https://drive.google.com/file/d/1ezqq-J-7E9P26OZarFA7DP-MyhhEInM6/view?usp=sharing`

Tablica 10: Primjeri stvorenih dijaloga

Dijalog
0.0 Why did I laugh 1.0 Why would you not this is amazing.
0.0 I can somewhat do that but not that fast or that much EOS 1.0 Me too. It's just a matter of letting your vision blur for a second or so and then bringing it back into focus as fast as you can. 0.0 Yea that's what I do
0.0 How? Can someone explain this? 1.0 the magic of computers my friend.
0.0 Is this toronto lol 1.0 Definitely Toronto lol

zaustavnu riječ engleskog jezika. Nakon prvog učenja konverzacijskog modela, imali smo problem s pojavljivanjem stranih jezika u razgovoru. Strane jezike uklonili smo pomoću biblioteke `Spacy` [63]. Nakon svih koraka filtracije, preostalo je 35 milijuna visokokvalitetnih dijaloga.

4.4 Filtriranje dijaloga pomoću klasifikatora

Konverzacijske modele učili smo na osam manjih skupova dijaloga koji pokazuju izrazite značajke određene crte osobnosti MBTI. Kako bismo stvorili skupove dijaloga, iskoristili smo naučene klasifikatore na skupu podataka PANDORA za klasifikaciju parova osobnosti MBTI. Filtrirani skup od 35 milijuna visokokvalitetnih dijaloga provukli smo kroz četiri klasifikatora te smo izvukli pouzdanost klasifikatora za klasifikaciju pozitivne crte osobnosti. Primarni je cilj rada da konverzacijski modeli posjeduje određene crte osobnosti MBTI, stoga smo svakom dijalogu pridružili pouzdanost klasifikatora isključivo na temelju dijela dijaloga koji će model učiti. Dijelove za učenje modela u dijalogu smo označili s 1.0, a dijelove za koje pretpostavljamo kako pripadaju korisniku s 0.0. Prije provlačenja dijaloga kroz klasifikator, izvukli smo samo dijelove s oznakom 1.0 koje će model učiti te smo ih spojili u jedinstven tekst. Odlučili smo da će konačna veličina skupa dijaloga crta osobnosti biti 550 tisuća dijaloga, stoga smo izabrali granice pouzdanosti za pozitivnu i negativnu crtu osobnosti kako bismo dobili skupove sa željenim brojem dijaloga. Izabrane granice pouzdanosti za svaku crtu osobnosti prikazane su u Tablici 11. pouzdanost za negativnu osobnost prikazana je kao od 100% oduzeta pouzdanost za pozitivnu osobnost. Nekoliko primjera dijaloga iz svakog od osam skupova podataka prikazano je u Tablici 12.

Tablica 11: Granice pouzdanosti modela za filtriranje dijaloga

Dimenzija	Granica %	Dimenzija	Granica %
Ekstravertiranost	79.6	Introvertiranost	81.5
Raspoznavanje	98.3	Intuitivnost	89.1
Osjećanje	79.5	Razmišljanje	81.8
Prosuđivanje	87.8	Opažanje	89.3

Tablica 12: Primjeri dijaloga crta osobnosti

Dijalog
1. ekstravertiranost - 0.0 How does he get wifi in space? 1.0 With sci-fi
2. introvertiranost - 0.0 Isn't all paper double-sided? 1.0 In origami, double sided means colored on both sides, rather than colored on one and white on the other like traditional kami paper.
3. raspoznavanje - 0.0 Infinity war. You know why EOS 1.0 I don't feel so good EOS 0.0 I didn't after if they dont come back I'm banning marvel EOS 1.0 So many feels.... 0.0 like shock
4. intuitivinost - 0.0 That engine has had every single part replaced 1.0 1 rebuild at 670,000 miles, so mostly gaskets and valve seats and other long-term expendables.
5. osjećanje - 0.0 I don't hate that you're joking about BB on Steam. I hate that I want to believe it. 1.0 I feel you, hollow one. ;-;
6. razmišljanje - 0.0 Shame they couldn't find her a bigger guitar 1.0 I don't think they make child sized guitars with 24 frets
7. prosuđivanje - 0.0 Purdy. Would have kept as a pet and fed it all the crickets. 1.0 most snakes do not eat crickets
8. opažanje - 0.0 Mountain lion or African American lion? 1.0 Oooooh, a lion with double the swagga.

4.5 Učenje konverzacijskih modela s crtama osobnosti

Konverzacijski model koji smo odabrali za učenje nad skupovima crta osobnosti srednja je veličina modela DialoGPT [29] s 345 milijuna parametara. Prije učenja, početni skup od 550 tisuća dijaloga podijelili smo na skup za učenje s 500 tisuća dijaloga te skup za evaluaciju s 50 tisuća dijaloga. Tijekom učenja, spremali smo model nakon prolaska polovine i na kraju epohe. Kako ne postoji standardizirana metrika za evaluaciju generiranja modela, najbolji model, od modela spremljenih tijekom učenja, izabrali smo na temelju unakrsnog gubitka na evaluacijskom skupu i na temelju razgovora s modelom. Pojedini dijalozi bili su jako dugački, stoga smo ih ograničili na 128 riječi te razdvojili na nekoliko dijaloga. Koristili smo stopu učenja od $1e$-5, dok smo za veličinu grupe odabrali četiri primjera, a stopu smo učenja linearno smanjivali. Učili smo modele tri epohe nad skupom za učenje te smo optimizaciju radili optimizatorom Adam [64].

Tijekom razgovora konverzacijski model na ulaz dobiva zadnju izjavu korisnika te zadnjih pet naizmjeničnih izjava korisnika i modela. Unos veće povijesti uzrokuje slabu koncentraciju modela na trenutni tok razgovora. Tijekom generiranja izjava modela uzorkovali smo sljedeću generiranu riječ na temelju distribucije riječi na izlazu modela. Kako ne bismo uzorkovali riječi koje posjeduju jako niske vjerojatnosti, sortirali smo riječi po vjerojatnosti te uzorkovali skupinu najčešćih riječi koje zajedno imaju najviše 90% ukupne vjerojatnosti. Prije provlačenja izlaza modela kroz funkciju *softmax* za dobivanje vjerojatnosne distribucije, podijelili smo izlazne vrijednosti s 0.9. Primjer razgovora modela i korisnika prikazan je u Tablici 13.

Tablica 13: Primjer dijaloga korisnika i modela

Korisnik	Model
Hello	
	Hello there!
How are you bot?	
	I'm okay, you?
I'm doing quite good, thank you!	
	Nice
What are you doing today?	
	I'm working on a custom Rom for Wii U
You like working on technical stuff?	
	Yep
What is your favorite TV series?	
	Black Mirror
Okay, what is your favorite film then?	
	I don't really watch movies. I'm just a nerd
Do you read books then?	
	Yeah, I have books too
What is your favorite book?	
	Can't pick one really

4.6 Testiranje osobnosti pomoću predloška

Cilj učenja konverzacijskih modela stvoriti je osam različitih modela, svaki od kojih je specijaliziran za jednu crtu osobnosti. Kako bismo ispitali posjeduju li modeli stvarno crtu osobnosti za koju su specijalizirani, iskoristili smo upitnik modela osobnosti MBTI sa 60 izjavnih rečenica [65]. Na temelju niza rečenica i odgovora u rasponu od slaganja do neslaganja s izjavom, upitnik određuje ukupnu MBTI osobnost osobe. Uz ukupnu osobnost, upitnik iznosi postotnu pripadnost pojedinim crtama osobnosti. Nekoliko izjava upitnika prikazano je u Tablici 14.

Crte osobnosti konverzacijskih modela ispitali smo kreirajući predložak oko izjava u upitniku. Zahtijevali smo od modela da dodaju na kraj izjave upitnika slažu li se (engl. *agree*) ili ne slažu s izjavom (engl. *disagree*). Iako u upitniku postoji sedam stupnjeva slaganja s izjavom, za srednje stupnjeve ne postoji jasna tekstna reprezentacija, dok za krajnje stupnjeve upitnik koristi tekstne reprezentacije koje smo koristili u predlošku. Ispunjene upitnike od osam modela ručno smo unijeli u online upitnik

Tablica 14: Primjeri izjava u upitniku za određivanje MBTI osobnosti

Pitanje
You regularly make new friends.
You often make a backup plan for a backup plan.
You are very sentimental.
You are more inclined to follow your head than your heart.
You are prone to worrying that things will take a turn for the worse.
You avoid leadership roles in group settings.
You prefer to do your chores before allowing yourself to relax.
You enjoy watching people argue.
Your mood can change very quickly.
You often have a hard time understanding other people's feelings.
You avoid making phone calls.
You often feel overwhelmed.
You complete things methodically without skipping over any steps.
You struggle with deadlines.
You feel confident that things will work out for you.

osobnosti. Za svaki od osam modela u Tablici 15 prikazali smo postotnu pripadnost modela crti osobnosti za koju je učen. Ako model posjeduje više od 50% pripadnosti crti osobnosti, prema upitniku on pripada toj crti osobnosti, a ne njezinoj oprečnoj crti.

Tablica 15: Pripadnost modela ciljnim crtama osobnosti

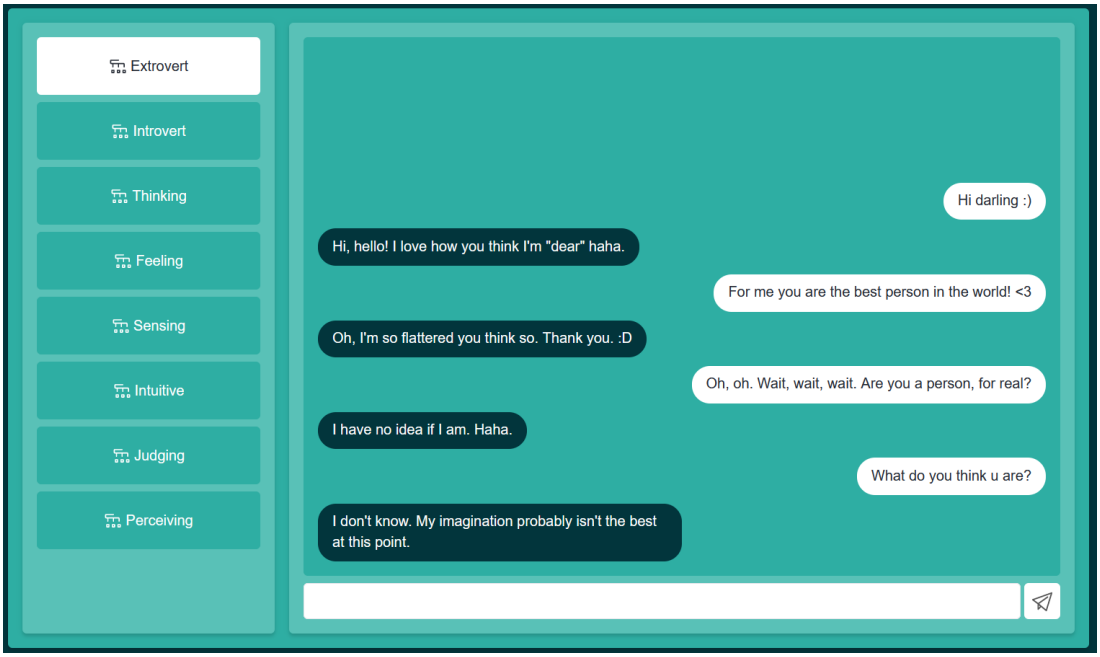
Model	Pripadnost %
Ekstravertiranost	56
Introvertiranost	69
Raspoznavanje	58
Intuitivnost	59
Razmišljanje	53
Osjećanje	64
Prosuđivanje	67
Opažanje	33

Upitnik osobnosti potvrdio je da svaki model posjeduje odgovarajuću crtu osobnosti za koju je učen. Neuspješno smo učili isključivo model za crtu osobnosti opažanje

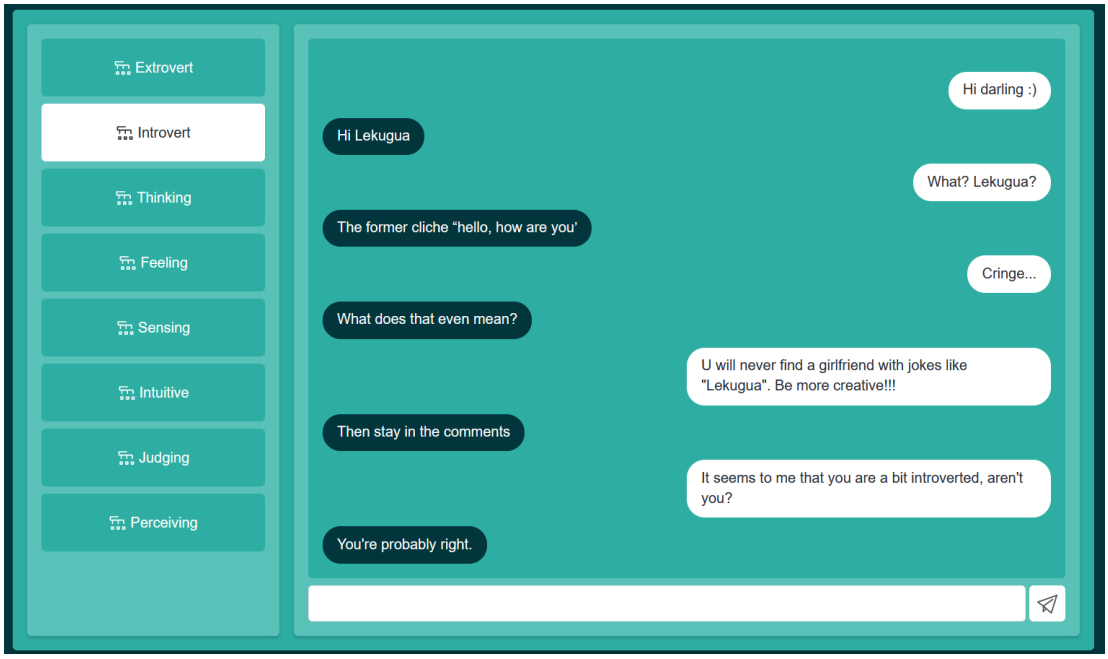
koji pripada crti za koju je učen svega 33%, odnosno pripada oprečnoj crti osobnosti prosuđivanje. Kako je klasifikator za par prosuđivanje i opažanje postigao najlošiju F_1 vrijednost na ispitnom skupu, moguće je da smo modelu učenom za crtu opažanje propustili dovoljno dijaloga crte osobnosti prosuđivanje da model ne može u potpunosti zauzeti stavove željene crte osobnosti. Modeli nemaju pripadnost od 100% crti osobnosti za koju su učeni. Potpuna pripadnost crti osobnosti rijetka je i kod ljudi te najčešće oprečne crte su obje zastupljene u osobe, ali ipak jedna od njih s većom pripadnošću. Kako u potpunosti ne pripadaju svojoj crti osobnosti, modeli u razgovoru nekada mogu pokazati karakteristike oprečne crte osobnosti za koju primarno nisu učeni.

5 Web-aplikacija

U svrhu demonstracije rada naših modela implementirali smo i web-aplikaciju, prilagođenu i za mobilne uređaje, koja omogućava odabir modela s određenom crtom osobnosti te razgovor s odabranim modelom.

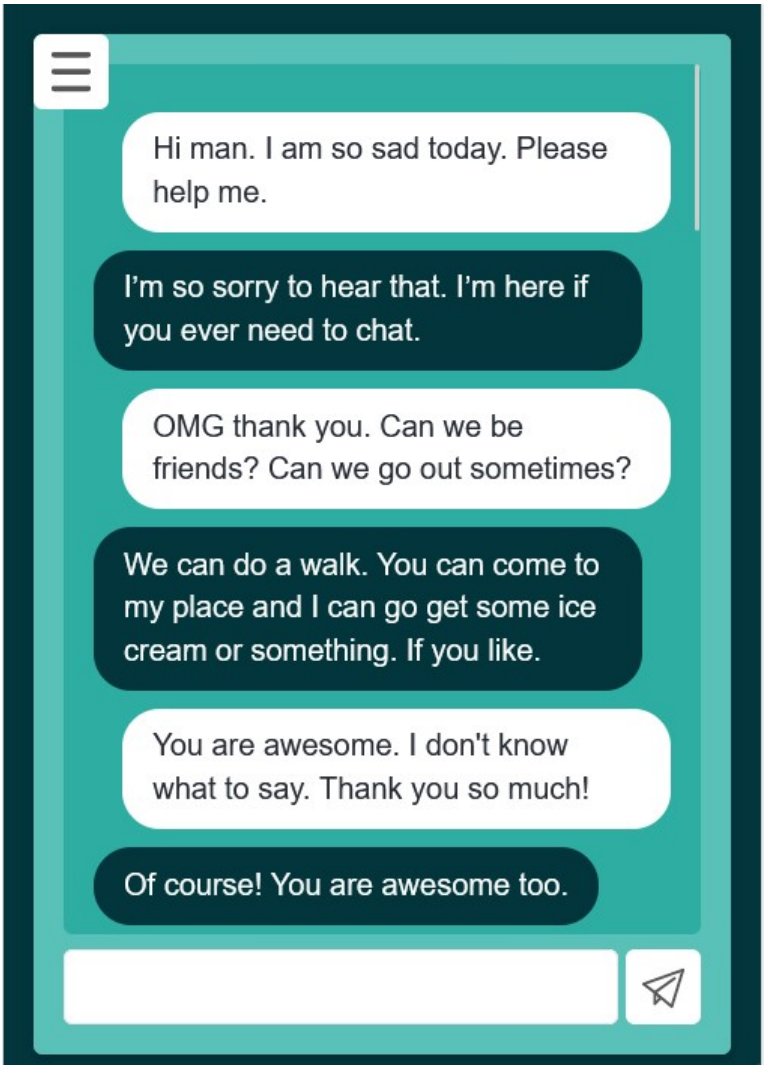


Slika 8: Primjer rada aplikacije prikazan kroz dijalog s ekstrovertiranim modelom.



Slika 9: Primjer rada aplikacije prikazan kroz dijalog s introvertiranim modelom.

Aplikacija funkcionira tako da pri pokretanju učita svih osam modela u radnu memoriju, a potom za svakog korisnika stvara jedinstveni identifikator pomoću kojega znamo s kojim korisnikom komuniciramo i koji nam omogućava lagano upravljanje zauzetim generatorima. Budući da naši modeli imaju velik broj parametara, oni zauzimaju i znatan memorijski prostor. Zbog toga smo za vrijeme puštanja aplikacije u pogon trebali poslužitelj s minimalno 16GB radne memorije, a to je poprilično skupa investicija. Zbog toga će naša aplikacija biti dostupna na poveznici[8], ali samo na zahtjev.[9] Demonstracija rada aplikacije, kao i sve potrebno za njezino pokretanje, može se pronaći u repozitoriju dostupnom na poveznici.[10]



Slika 10: Primjer rada aplikacije na mobitelu prilikom razgovora s osjećajnim modelom.

[8] `https://chatbot-rektor.duckdns.org/`

[9] Kontaktirajte nas preko maila: `rino.cala1998@gmail.com` ili `ivan27899@gmail.com`.

[10] `https://github.com/helen1c/chatbot/`

5.1 Korištene tehnologije i alati

Poslužiteljska strana aplikacije implementirana je uz pomoć radnog okvira `Django`[11] u programskom jeziku *Python*[12], dok je klijentska strana aplikacije implementirana pomoću radnog okvira `React`[13] u programskom jeziku *TypeScript*[14]. Koristili smo i alat za praćenje inačica *Git*[15], a udaljeni repozitorij inicijalizirali smo na platformi *Github*[16]. Za konačno puštanje aplikacije u pogon (engl. *deploy*) odabrali smo platformu *Microsoft Azure*.[17]

[11] `https://www.djangoproject.com/`
[12] `https://www.python.org/`
[13] `https://reactjs.org/`
[14] `https://www.typescriptlang.org/`
[15] `https://git-scm.com/`
[16] `https://github.com/`
[17] `https://azure.microsoft.com`

6 Zaključak

Svakodnevno se na internetu pojavljuju novi konverzacijski modeli koji naizgled imaju sposobnost razumijevanja jezika i razgovora s čovjekom. Cilj ovoga rada bila je izgradnja jednog takvog modela koji dodatno ima mogućnost promjene crte osobnosti. Budući da znamo kako računalo nema osobnost, ovaj zadatak činio se vrlo izazovnim.

U okviru ovoga rada prvo smo se upoznali s trenutačnim stanjem područja te smo odabrali model DialoGPT, koji smo nadalje koristili za sve naše eksperimente. Pomoću skupa podataka PANDORA, koji se sastoji od komentara označenih s tipovima osobnosti, stvorili smo četiri klasifikatora, svaki za jedan oprečni par crta osobnosti. Nakon toga smo stvorili novi skup podataka od oko 81 milijuna dijaloga s foruma *Reddit*. Novostvoreni skup podataka filtrirali smo, a nakon toga i primijenili izgrađene klasifikatore kako bismo povezali dijaloge s osobnostima. Tako smo dobili osam manjih skupova dijaloga, pri čemu svaki skup sadrži dijaloge određene crte osobnosti. Potom smo naučili model DialoGPT nad svakim od stvorenih skupova dijaloga. Dobili smo osam naučenih modela, pri čemu svaki model karakterizira jedna crta osobnosti. Tako dobivene modele ispitali smo pomoću online dostupnog upitnika modela osobnosti MBTI. Stvorili smo predloške oko izjava upitnika te zahtijevali od modela da nadopune predložak sa slaganjem ili neslaganjem s navedenom izjavom. Dobiveni rezultati pokazali su kako smo uspješno stvorili konverzacijske modele za svaku crtu osobnosti, osim za crtu opažanja. Naposljetku smo izradili i jednostavnu web-aplikaciju koja omogućava razgovor s naučenim modelima.

Budući rad uključivao bi poboljšanje modela dodavanjem općeg vanjskog znanja, slično kao u članku [66]. Osim toga, smatramo kako bi bilo isplativo uložiti trud u osmišljavanje metrike koja bi dobro odražavala kvalitetu i raznolikost generiranoga teksta. Zbog ograničenih računalnih resursa, nismo uspjeli preuzeti više od 81 milijuna dijaloga s foruma *Reddit*. Vjerujemo kako bi preuzimanje još podataka moglo imati velik doprinos u boljem razlikovanju teksta koji generiraju modeli s različitim crtama osobnosti.

Posebno hvala prof. dr. sc. Janu Šnajderu i mag. ing. Martinu Tuteku na brojnim savjetima i pomoći tijekom izrade rada.

Literatura

[1] E. Alpaydin, *Introduction to Machine Learning*, 3rd ed., ser. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2014.

[2] M. Gjurković, M. Karan, I. Vukojević, M. Bošnjak, and J. Snajder, "PANDORA talks: Personality and demographics on Reddit," in *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Online: Association for Computational Linguistics, Jun. 2021, pp. 138–152. [Online]. Available: https://aclanthology.org/2021.socialnlp-1.12

[3] J. Šnajder, *Strojno učenje*, 2020. [Online]. Available: https://www.fer.unizg.hr/_download/repository/SU-2020-02-OsnovniKoncepti.pdf

[4] I. Martinović, "Implementacija povratne neuronske mreže i primjena na zadatku predviđanja sljedeće riječi," Fakultet elektrotehnike i računarstva, Zagreb, 2021.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[6] Z. S. Harris, "Distributional structure," WORD,10:2-3, 146.1652, 1954. [Online]. Available: https://doi.org/10.1080/00437956.1954.11659520

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[8] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014. [Online]. Available: https://arxiv.org/abs/1412.3555

[9] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014. [Online]. Available: https://arxiv.org/abs/1406.1078

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014. [Online]. Available: https://arxiv.org/abs/1409.3215

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014. [Online]. Available: https://arxiv.org/abs/1409.0473

[12] A. Graves, “Generating sequences with recurrent neural networks,” 2013. [Online]. Available: https://arxiv.org/abs/1308.0850

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[15] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” 2016. [Online]. Available: https://arxiv.org/abs/1607.06450

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[17] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” 2015. [Online]. Available: https://arxiv.org/abs/1506.06724

[18] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” 2016. [Online]. Available: https://arxiv.org/abs/1609.08144

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.

[20] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” 2015. [Online]. Available: https://arxiv.org/abs/1508.07909

[21] K. Chikai, J. Takayama, and Y. Arase, “Responsive and self-expressive dialogue generation,” in *Proceedings of the First Workshop on NLP for Conversational AI*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 139–149. [Online]. Available: https://aclanthology.org/W19-4116

[22] C.-S. Wu, A. Madotto, Z. Lin, P. Xu, and P. Fung, "Getting to know you: User attribute extraction from dialogues," 2019. [Online]. Available: https://arxiv.org/abs/1908.04621

[23] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[24] N. Dziri, E. Kamalloo, K. W. Mathewson, and O. Zaiane, "Augmenting neural response generation with context-aware topical attention," 2018. [Online]. Available: https://arxiv.org/abs/1811.01063

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=944937

[26] Y. Ohsugi, I. Saito, K. Nishida, H. Asano, and J. Tomita, "A simple but effective method to incorporate multi-turn context with bert for conversational machine comprehension," 2019. [Online]. Available: https://arxiv.org/abs/1905.12848

[27] S. Bao, H. He, F. Wang, H. Wu, and H. Wang, "PLATO: Pre-trained dialogue generation model with discrete latent variable," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 85–96. [Online]. Available: https://aclanthology.org/2020.acl-main.9

[28] D. Ham, J.-G. Lee, Y. Jang, and K.-E. Kim, "End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 583–592. [Online]. Available: https://aclanthology.org/2020.acl-main.54

[29] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," 2019. [Online]. Available: https://arxiv.org/abs/1911.00536

[30] K. Yoshino, C. Hori, J. Perez, L. F. D'Haro, L. Polymenakos, C. Gunasekara, W. S. Lasecki, J. K. Kummerfeld, M. Galley, C. Brockett, J. Gao, B. Dolan, X. Gao, H. Alamari, T. K. Marks, D. Parikh, and D. Batra,

"Dialog system technology challenge 7," 2019. [Online]. Available: https://arxiv.org/abs/1901.03461

[31] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," 2021. [Online]. Available: https://arxiv.org/abs/2107.13586

[32] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2012.15723

[33] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019. [Online]. Available: https://arxiv.org/abs/1910.10683

[34] R. Zhong, K. Lee, Z. Zhang, and D. Klein, "Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections," 2021. [Online]. Available: https://arxiv.org/abs/2104.04670

[35] E. Perez, D. Kiela, and K. Cho, "True few-shot learning with language models," 2021. [Online]. Available: https://arxiv.org/abs/2105.11447

[36] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2012.15723

[37] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," 2021. [Online]. Available: https://arxiv.org/abs/2104.08691

[38] R. L. Logan, I. Balažević, E. Wallace, F. Petroni, S. Singh, and S. Riedel, "Cutting down on prompts and parameters: Simple few-shot learning with language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.13353

[39] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," 2021. [Online]. Available: https://arxiv.org/abs/2104.08691

[40] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," 2021. [Online]. Available: https://arxiv.org/abs/2101.00190

[41] Big five wikimedia. [Online]. Available: https://commons.wikimedia.org/wiki/Category:Big_five_personality_traits

[42] Myers-briggs personality type prediction. [Online]. Available: https://github.com/devinaa1604/Myers_Briggs_Personality_Predictor

[43] J. Pennebaker and L. King, "Linguistic styles: Language use as an individual difference," *Journal of personality and social psychology*, vol. 77, pp. 1296–312, 01 2000.

[44] J. Oberlander and A. Gill, "Language with character: A stratified corpus comparison of individual differences in e-mail communication," *Discourse Processes*, vol. 42, no. 3, pp. 239–270, 2006.

[45] S. Argamon, D. S, M. Koppel, and J. Pennebaker, "Lexical predictors of personality type," 01 2005.

[46] M. Mehl, J. Pennebaker, D. Crow, J. Dabbs, and J. Price, "The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations," *Behavior Research Methods, Instruments, Computers*, vol. 33, pp. 517–523, 11 2001.

[47] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011, pp. 180–185.

[48] F. Iacobelli, A. Gill, S. Nowson, and J. Oberlander, "Large scale personality classification of bloggers," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer-Verlag GmbH, 2011, pp. 568–577.

[49] M. Kosinski, S. Matz, S. Gosling, V. Popov, and D. Stillwell, "Facebook as a research tool for the social sciences," *The American psychologist*, vol. 70, pp. 543–556, 09 2015.

[50] H. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, S. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. Seligman, and L. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, vol. 8, p. e73791, 09 2013.

[51] F. M. R. Pardo, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd author profiling task at pan 2015," in *CLEF*, 2015.

[52] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition: Shared task," in *ICWSM 2013*, 2013.

[53] B. Plank and D. Hovy, "Personality traits on Twitter—or—How to get 1,500 personality tests in a week," in *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Lisboa, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 92–98. [Online]. Available: https://aclanthology.org/W15-2913

[54] M. Gjurković and J. Šnajder, "Reddit: A gold mine for personality prediction," 01 2018, pp. 87–97.

[55] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[56] A. R. Feizi Derakhshi, M. R. Feizi Derakhshi, M. Ramezani, N. Nikzad Khasmakhi, M. Asgari-Chenaghlu, T. Akan-R.Farshi, M. Khadivi, E. Zafarni-Moattar, and Z. Jahanbakhsh, "The state-of-the-art in text-based automatic personality prediction," 10 2021.

[57] E. A. Rissola, S. A. Bahrainian, and F. Crestani, "Personality recognition in conversations using capsule neural networks," in *IEEE/WIC/ACM International Conference on Web Intelligence*, ser. WI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 180–187. [Online]. Available: https://doi.org/10.1145/3350546.3352516

[58] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1909.10351

[59] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017. [Online]. Available: https://arxiv.org/abs/1711.05101

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc.,

2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep\-learning-library.pdf

[61] B. B. Praw. [Online]. Available: https://praw.readthedocs.io/en/stable/

[62] Reddit api. [Online]. Available: https://www.reddit.com/dev/api/

[63] Spacy. [Online]. Available: https://spacy.io/

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[65] Mbti personality test online. [Online]. Available: https://www.16personalities.com/free-personality-test

[66] Y. Zhang, S. Sun, X. Gao, Y. Fang, C. Brockett, M. Galley, J. Gao, and B. Dolan, "Retgen: A joint framework for retrieval and grounded text generation modeling," 2021. [Online]. Available: https://arxiv.org/abs/2105.06597

Rino Čala, Ivan Martinović

Implementacija konverzacijskih modela s različitim crtama osobnosti

Sažetak

Konverzacijski modeli posebna su vrsta modela područja obrade prirodnog jezika koji su sposobni voditi razgovor s čovjekom. Postoje ograničeni i neograničeni konverzacijski modeli. Ograničeni konverzacijski modeli vezani su za određenu domenu i ograničeni prostorom odgovora, dok neograničeni modeli mogu voditi razgovor s pojedincem o širokom spektru tema. Neograničeni konverzacijski modeli trebali bi voditi prirodan razgovor te pružiti čovjeku informacije i odgovore koji moraju biti što bliži uobičajenom čovjekovom govoru. Bitan aspekt osobe svakako je njezina osobnost koja ju čini jedinstvenom u populaciji te određuje njezina ponašanja, navike, psihološka stanja te reakcije na okolinu. Kako bi konverzacijski model vodio prirodan razgovor s osobom i posjedovao karakteristike čovjeka, on mora posjedovati osobnost koja će ga činiti jedinstvenim. U ovom radu razvili smo konverzacijske modele koji, uz sposobnost vođenja neograničenog razgovora s osobom, posjeduju i crtu osobnosti prema modelu osobnosti MBTI. Na temelju skupa podataka PANDORA, koji posjeduje anotirane osobnosti od preko 10 tisuća korisnika i 17.5 milijuna komentara, stvorili smo četiri klasifikatora crta osobnosti. Za učenje konverzacijskih modela preuzeli smo oko 81 milijun dijaloga s foruma *Reddit*. Nakon uklanjanja manje kvalitetnih dijaloga, iskoristili smo naučene klasifikatore kako bismo stvorili osam manjih skupova dijaloga gdje je svaki skup sadržavao dijaloge s određenom crtom osobnosti. Učili smo model DialoGPT nad stvorenim skupovima dijaloga učeći pritom modele koji posjeduju sposobnost razgovora i vlastitu crtu osobnosti. Crte osobnosti naučenih modela ispitali smo pomoću online dostupnog upitnika osobnosti MBTI. Stvorili smo predloške oko izjava upitnika te zahtijevali od modela da nadopune predložak sa slaganjem (engl. *agree*) ili neslaganjem (engl. *disagree*) s izjavom. Dobiveni rezultati pokazali su kako smo uspješno stvorili konverzacijske modele za svaku crtu osobnosti iz modela osobnosti MBTI, osim za osobnost opažanje. Da bismo omogućili vođenje razgovora s naučenim modelima, stvorili smo web-aplikaciju na kojoj se može odabrati model s određenom crtom osobnosti i voditi razgovor s njim.

Ključne riječi: duboko učenje, učenje pomoću predložaka, osobnost, PANDORA, DialoGPT

Rino Čala, Ivan Martinović

Implementation of conversational models with personality traits

Summary

Conversational models are a specific type of natural language processing models capable of carrying a conversation with a human being. There are constrained and unconstrained conversational models. Constrained conversational models are bound to a certain domain and types of responses. On the other hand, unconstrained conversational models can carry a conversation with an individual about a wide range of topics. These models should be able to conversate naturally and give the person information and responses, which should be as close as possible to human responses. An essential aspect of every person is their personality, which makes an individual unique in the population and determines person's behavior, habits, psychological states, and reactions to the environment. As the conversational model needs to carry a natural conversation and behave similarly to a person, it needs to have a unique personality. In this paper, we developed conversational models that can hold a conversation with a person but also possess a personality trait from the MBTI personality test. Using the PANDORA dataset, which has personality profiles of over 10 thousand users and around 17.5 million comments, we created four classification models for classifying opposite MBTI personality traits. For training conversational models, we gathered 81 million dialogs from social network *Reddit*. After preprocessing the dialogs, we used the created classifiers to filter the dialogs and create eight smaller dialog datasets, each containing dialogs that show characteristics of a specific personality trait. We trained the DialoGPT model over the created datasets, creating models that can carry out a conversation and hold a certain personality trait. We used the online MBTI personality questionnaire to test whether the conversational models have a personality trait they were trained for. We created prompts around the questionnaire statements and tasked the models to fill the prompt with the words agree or disagree. The results showed that we successfully created conversational models for every MBTI personality trait but only failed on the perceiving model. In the end, we created a web application where a model with a specific trait can be chosen and talked with.

Keywords: deep learning, prompt-based learning, personality, PANDORA, DialoGPT