

Chatbot u oblasti medicine baziran na Enkoder-Dekoder arhitekturi

Teodora Maruna

Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
teodora.maruna@gmail.com

Vladimir Jovanović

Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
jovanovicvladimir099@gmail.com

Apstrakt — Savremeni čovek je suočen sa sve većim rizikom od ugroženog zdravstvenog stanja zbog nepovoljnih životnih navika, što značajno može uticati na kvalitet života. Kada se pojave zdravstveni problemi, jedini trenutno adekvatan i siguran način za rešavanje je konsultacija sa lekarom. Međutim, postavlja se pitanje šta se dešava kada lekari, usled velikog broja pacijenata, nemaju dovoljno vremena da pruže adekvatan savet i negu svakom pojedincu. Kao odgovor za navedeni problem, predlažemo kreiranje konverzijskog *chatbot*-a koji koristi napredne algoritme mašinskog učenja i tehnike procesiranja prirodnog jezika kako bi pružio preporuke pacijentima pre konsultacija sa lekarom. S obzirom da su rekurentni enkoder-dekoder modeli dominantni u oblasti modelovanja konverzijskih dijaloga sistema, za implementaciju *chatbot*-a u ovom radu, koristi se *sequence to sequence* model, koji poseduje enkoder-dekoder arhitekturu i ima implementiran mehanizam pažnje (eng. *attention mechanism*) između enkoder i dekoder sloja. U okviru rada izneti su eksperimentalni rezultati koji pokazuju da su performanse bolje u slučaju kad se *chatbot* obučava tako da za slična pitanja dobija generalizovane, opšte odgovore, a ne odgovore koji su vrlo usko specijalizovani za navedeno pitanje pacijenta. Prilikom pregleda srodnih radova, primećeno je da dolazi do problema nedostaka memorije. Kako bi se prevazišao navedeni problem, predložena je mogućnost upotrebe inkrementalnog obučavanja, pri čemu su iznete vrednosti BLEU metrike koje pokazuju da je moguće sprovesti postupak inkrementalnog obučavanja sa prevazilaženjem problema gubitka stečenog znanja. Kreiranjem *chatbot*-a, otvaraju se nove mogućnosti za unapređenje zdravstvenog sistema.

Ključne reči — *seq2seq*, *enkoder-dekoder*, *NLP*, *mehanizam pažnje*, *chatbot*

I. UVOD I RELEVATNA LITERATURA

Opšte zdravstveno stanje svake osobe je jedan od glavnih faktora kvalitetnog života. Kada je zdravstveno stanje ugroženo, prva stanica kojoj se obraćamo je lekar. Kako bi lekar pružio neophodnu pomoć pacijentima, on mora da izdvoji određeno vreme, a vreme svakog lekara je ograničeno i dragoceno. Zbog toga, održavanje konsultacija sa lekarom za svaki zdravstveni problem je praktično nemoguće. To dovodi do ideje kreiranja *chatbot*-a koji će koristiti napredne AI algoritme da postavi osnovnu dijagnozu pacijenta i da mu da preporuke pre konsultacija sa lekarom. *Chatbot* bi imao kao cilj da pomogne u smanjenju troškova zdravstvene zaštite ali bi istovremeno omogućio i veću pristupačnost medicinskom znanju.

U ovom radu, opisan je postupak razvoja konverzijskog (eng. *conversational*) *chatbot*-a za predikciju dijagnoze i davanje preporuka za lečenje pacijenata na osnovu simptoma koje pacijent navede [1]. Konverzijski *chatbot* je vrsta *chatbot*-a koji simulira razgovor sa korisnikom i njegova zamisao je shvatanje namere korisnika i davanje adekvatnog odgovora na osnovu analiziranih podataka. Konkretna predikcija dijagnoze se vrši na osnovu dokumentovanih razgovora između pacijenata i lekara.

Postoji više kategorija *chatbot*-ova, zavisno od algoritama na kojima su zasnovani. S jedne strane postoje *chatbot*-ovi zasnovani na *if-else* pravilima (eng. *rule-based chatbot*) i oni su ograničenih mogućnosti, ali postoje i oni koji koriste napredne algoritme mašinskog učenja i samim tim nude širi dijapazon mogućnosti. Ovaj rad se bavi opisom implementacije *chatbot*-a zasnovanog na algoritmima mašinskog učenja. Za implementaciju ovakih *chatbot*-ova, najčešće se koriste rekurentne neuronske mreže (eng. *Recurrent Neural Network*) [2]. Rekurentne neuronske mreže imaju mogućnost pamćenja prethodnog stanja i zato se koriste za rešavanje problema predikcije sekvenci.

U nastavku će biti navedeni radovi srodnih istraživanja.

Jedan takav rad je 'Seq2Seq AI Chatbot with Attention Mechanism' [3]. Cilj ovog projekta je razvoj konverzijskog *chatbot*-a za opšte namene. Skup podataka korišten za obučavanje je *Cornell Movie Dialog Corpus Dataset*¹, koji predstavlja jedan od najčešće korišćenih skupova podataka za obučavanje *chatbot*-ova [4]. Iz skupa podataka izbačene su sve rečenice koje sadrže više od 5 reči, čime je pojednostavljeno obučavanje. Istovremeno je zaobiđen problem nedostatka memorije do kojeg dolazi prilikom pokušaja obučavanja mreže sa ogromnim skupom podataka na prosečnom personalnom računaru. *Chatbot* je kreiran primenom enkoder-dekoder arhitekture [5]. *Bahdanau* mehanizam pažnje (eng. *attention mechanism*) je implementiran između enkoder i dekoder slojeva i predstavlja tehniku koja omogućava modelu da se fokusira na relevantne delove ulaznih podataka tokom obrade sekvenci [6]. U okviru rada, poređene su 3 konfiguracije sa različitom kombinacijom hiperparametara, a za određivanje najboljeg modela korištena je ljudska estimacija. Opisana arhitektura je korištena prilikom

¹ https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

implementacije našeg *chatbot*-a. Koristili smo seq2seq (*Sequence To Sequence*) model, odnosno enkoder-dekoder arhitekturu [5]. Posmatrani rad ne rešava problem nedostatka memorije nego smanjuje obim skupa podataka za obučavanje. U okviru naše implementacije *chatbot*-a, u želji da rešimo problem nedostatka memorije, pokušali smo da implementiramo inkrementalno obučavanje (eng. *incremental learning*).

‘An Empathetic Conversational Agent with Attentional Mechanism’ je rad u kome su predstavljena i upoređena dva modela empatičnih *chatbot*-ova [7]. Oba modela *chatbot*-a predstavljaju seq2seq modele, ali jedan je implementiran dodatno sa mehanizmom pažnje. Da bi mogao da generiše empatične odgovore, *chatbot* je obučen nad *Facebook AI Empathetic Dialogue*² skupom podataka. Za evaluaciju rezultata korištena je BLEU (*Bilingual Evaluation Understudy*) metrika [8]. Model *chatbot*-a sa mehanizmom pažnje imao je vrednost BLEU metrike 0.137 i pokazao je bolji rezultat od klasičnog seq2seq modela, čija je vrednost BLEU metrike 0.126. Model sa mehanizmom pažnje bolje razume suštinu i emociju korisnika izvorne rečenice i daje odgovor sa više empatije. S obzirom da korišćenje mehanizma pažnje utiče pozitivno na performanse *chatbot*-a, u našem radu je takođe implementiran mehanizam pažnje koji omogućava selektivno obraćanje pažnje na delove rečenica koje su od većeg značaja.

Rad opisan u ‘Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic’ koristi NLP (*Natural Language Processing*) tehnike i fuzzy SVM (*Support Vector Machine*) klasifikator za razvoj *chatbot*-a za predviđanje dijagnoze pacijenta na osnovu unetih simptoma [9], [10], [11]. Napre se vrši procesiranje teksta NLP tehnikama - uklanjanje nerelevantnih podataka, tokenizacija, parsiranje. Zatim se vrši obučavanje SVM klasifikatora na osnovu prethodno procesiranih podataka NLP tehnikama. Za evaluaciju performansi sistema korišćena je BLEU metrika i dobijena vrednost BLEU metrike je 0.2529. NLP tehnike spomenute u ovom naučnom radu, iskorištene su i u okviru našeg rada.

Ideja rada ‘Task-oriented Dialogue System for Automatic Diagnosis’ je kreiranje *chatbot*-a za automatsku dijagnozu pacijenata na osnovu unesenih simptoma [12]. Prilikom konverzacije sa pacijentom, sistem je osposobljen da postavlja pitanja o propratnim simptomima kako bi bila izvršena tačna dijagnoza. Ovo je omogućeno korišćenjem učenja sa potkrepljenjem (eng. *reinforcement learning*) [13]. Autori projekta su napravili svoj skup podataka koji se sastoji od dve vrste podataka. Prvi skup podataka su podaci koje pacijent u toku pregleda inicijalno iznosi lekaru, dok drugi skup podataka predstavlja one simptome koje doktor dobija nakon dalje konverzacije sa pacijentom, ispitujući ga za propratne simptome. Prvi skup simptoma autori ovog rada nazivaju eksplicitnim simptomima, dok drugi skup podataka nazivaju implicitnim simptomima. Rezultati pokazuju da postojanje implicitnih simptoma povećava tačnost pri dijagnozi bolesti za više od 10%. To govori da se korišćenjem učenja sa

potkrepljenjem može izvršiti preciznija dijagnostika bolesti. S obzirom da je ista tematika navedenog naučnog rada i našeg projekta, odradili smo dublju analizu načina na koji je korišćen ovaj skup podataka. Došli smo do zaključka da implicitno prikupljanje simptoma i korišćenje učenja sa potkrepljenjem znatno poboljšavaju performanse. Metodologija predstavljena u ovom radu ostaje kao ideja za buduću nadogradnju našeg rada.

U poglavlju 2 opisan je proces pretprocesiranja i integracije skupa podataka, dok je u poglavlju 3 navedena metodologija korišćena za implementaciju *chatbot*-a. U poglavlju 4 prikazani su rezultati u okviru različitih konfiguracija, odnosno primenom različitih hiperparametara. Pored toga, poglavlje 4 se bavi diskusijom o određenim ograničenjima ovog projekta kao i poređenjem dobijenih rezultata sa rezultatima sličnih radova. U poglavlju 5 izveden je zaključak čitavog rada, a predloženi su i mogući pravci daljeg razvoja.

II. SKUP PODATAKA

Ovo poglavlje opisuje postupak kreiranja skupova podataka koji se koriste za obučavanje *chatbot*-a. Nad pronađenim podacima izvršene su odgovarajuće tehnike pretprocesiranja kako bismo uklonili nepotrebne delove i redundantne podatke.

U toku implementacije ovog rada kreirali smo dva skupa podataka i ovi skupovi se koriste nezavisno jedan od drugog zbog prirode sadržanih podataka. Priroda podataka dodatno je pojašnjena u narednim paragrafima. Zbog nemogućnosti integracije podataka u jedinstveni korpus usled različite prirode skupova podataka, u nastavku rada ćemo posebno posmatrati modele iste arhitekture ali obučene nad različitim skupovima podataka.

A. Skup podataka *MedDialogPlus*

U okviru naučnog rada ‘MedDialog: Two Large-scale Medical Dialogue Datasets’ naveden je skup podataka *MedDialog*³ koji sadrži dijaloge na kineskom i na engleskom jeziku [14]. Pošto ćemo se u našem radu fokusirati na engleski jezik, iskoristićemo samo deo korpusa koji se bavi engleskim jezikom. Engleski deo korpusa sadrži 514,908 iskaza od čega je 257,454 od strane pacijenata i isti toliki broj od strane lekara. Svaki dijalog se sastoji od dela u kome pacijent objašnjava svoje zdravstveno stanje i dela gde lekar daje odgovor. Podaci su preuzeti sa dva sajta - *iCliniq* i *HealthcareMagic*, koji predstavljaju online platforme za davanje stručnih medicinskih saveta.

Još jedan izvor podataka⁴, formiran je na osnovu podataka sa sledećih veb sajtova - *eHealth Forum*, *Question Doctors* i *WebMD*.

Skup podataka *MedDialogPlus* predstavlja integraciju dva navedena izvora podataka. S obzirom da su konverzacije pacijenata i lekara preuzete sa raznih online foruma, kod ovog skupa podataka za slična pitanja postoje različiti odgovori, jer

² <https://www.kaggle.com/datasets/atharvjairath/empathetic-dialogues-facebook-ai>

³ <https://drive.google.com/drive/folders/1g29ssimdZ6JzTST6Y8g6h-ogUNReBUD>

⁴ <https://github.com/LasseRegin/medical-question-answer-data>

je svaki odgovor lekara specijalno namenjen problemu koji je pacijent naveo. Odgovori su usko specijalizovani za navedeno pitanje pacijenta i njegove individualne probleme i ne predstavljaju uopšteno rešenje za datu tematiku pitanja.

B. Skup podataka *Intents*

*Intents*⁵ skup podataka se nalazi u json formatu. Svaki json objekat se sastoji iz 4-6 sličnih pitanja i jednog odgovora lekara koji odgovara za sva navedena pitanja. Skup podataka je obrađen tako da se struktura podudara sa strukturom realne konverzacije, gde se svaka konverzacija sastoji iz jednog pitanja pacijenta i adekvatnog odgovora lekara. *Intents* skup podataka sadrži odgovore koji su dovoljno generalizovani, uopšteni, nedvosmisleni i zbog toga uvek pokrivaju tematiku pitanja.

C. Tehnike pretprocesiranja

Skupovi podataka za oba modela *chatbot*-a su pretprocesirani na isti način. Proces pretprocesiranja se sastoji od sledećih koraka:

- (Korak I – Učitavanje i integracija) učitavanje i integracija json fajlova kako bi se podaci dalje mogli koristiti;
- (Korak II - Sumarizacija) sumarizacija pitanja pacijenata i odgovora lekara, za one rečenice koje su predugačke. Sumarizacija je implementirana pomoću *spacy*⁶ biblioteke. Maksimalna dužina pitanja pacijenta svedena je na 21 reč, a maksimalna dužina odgovora lekara svedena je na 158 reči. S obzirom da enkoder-dekoder arhitektura zahteva da je svaka sekvenca iste, fiksne dužine, sumarizacija je bila neophodna;
- (Korak III – Ekstrahovanje pitanja i odgovora) izdvajanje pitanja pacijenata u jednu listu, koja će biti ulaz za enkoder i odgovora lekara u drugu listu, koja će biti ulaz za dekoder. Ovo je odgovarajući format za enkoder-dekoder arhitekturu;
- (Korak IV - Prečišćavanje) prečišćavanja podataka na tako da su najpre uklonjeni nepotrebni karakteri i promenjen je format reči (i'm => i am...). Sve reči su prebačene na mala slova i uklonjeni su znakovi interpunkcije;
- (Korak V - Lematizacija) lematizacija (eng. *lemmatization*) predstavlja postupak svođenja reči na njen osnovni oblik. Izvršena je samo nad postavljenih pitanjima pacijenata, kako se ne bi izgubilo semantičko značenje odgovora. Lematizacija poboljšava tačnost jer omogućava *chatbot*-u da bolje identifikuje značenje reči i umanjuje konfuziju nastalu usled nailaska na različite formate iste reči. Pored toga, svođenjem reči na njen osnovni oblik smanjujemo obim vokabulara i doprinosimo boljim performansama *chatbot*-a;
- (Korak VI – Uklanjanje stop reči) uklanjanje stop reči je efikasna tehnika koja poboljšava performanse *chatbot*-a,

omogućavajući mu da se fokusira na bitne reči i njihov kontekst;

- (Korak VII – Kreiranje rečnika) za svaku reč u rečniku, definisana je konkretna reč i indeks reči. Usled velikog obima rečnika, rečnik je sveden na reči koje se ponavljaju 3 ili više puta;
- (Korak VIII - Tokenizacija) tokenizacija (eng. *tokenization*) predstavlja proces zamene svih reči u okviru pitanja pacijenata i odgovora lekara indeksima tih reči iz rečnika. Tokenizacija je neophodan korak jer ulaz u neuronsku mrežu mora da bude vektor, a ne niz karaktera. Pored toga, uvedeni su dodatni tokeni ('<PAD>', '<EOS>', '<OUT>', '<SOS>'). Token <PAD> koristi se za dodavanje tokena sekvenci tako da ona bude željene dužine. Token <OUT> označava reč koja se ne nalazi u rečniku. Token <SOS> predstavlja početak rečenice, dok token <EOS> označava kraj rečenice. Odrađeno je označavanja teksta koja ulazi u dekoder (odgovora lekara) tokenima <SOS> i <EOS>;
- (Korak IX - *Padding*) *padding* je tehnika kojom se dopunjuju kraće sekvence, dodavanjem specijalnih tokena na kraj sekvenci, da bi sve sekvence bile jednake dužine. Enkoder-dekoder arhitektura zahteva da svi ulazi u enkoder, odnosno u dekoder sloj budu fiksne dužine. Svim sekvencama koje ulaze u enkoder dodato je onoliko specijalnih tokena tako da ukupna dužina svake sekvence bude jednaka dužini najduže sekvence koja ulazi u enkoder. Analogno je urađeno i za sekvence koje ulaze u dekoder sloj.

D. Tehnika za reprezentovanje reči u vektorskom obliku

Word embedding je tehnika za predstavljanje reči u numeričkom obliku radi efikasnijeg rukovanja i analiziranja tekstualnih podataka. *Word embedding* uzima u obzir semantičku povezanost između reči. Konkretna *word embedding* koji je korišten pri implementaciji našeg *chatbot*-a je GloVe (*Global Vectors for Word Representation*) [15]. GloVe je algoritam mašinskog učenja koji se koristi za dobijanje vektorske reprezentacije reči. Razvijen je na Stanfordu 2014. godine. GloVe se oslanja na matricu brojeva međusobnog zajedničkog pojavljivanja reči (eng. *co-occurrence matrix*). Ta matrica se formira za okolinu reči. U početku se napravi velika matrica, nad kojom se vrši faktORIZACIJA. Ideja je da se smanji dimenzionalnost tako što će se zadržati komponente koji će u sebi sadržati većinu varijanse iz početnog skupa podataka.

III. METODOLOGIJA

Chatbot je vrsta dijalog sistem koji omogućava razgovor sa korisnikom, koristeći princip zasnovan na mehanizmu pitanje/odgovor. Zadatak *chatbot*-a predstavlja predikciju adekvatnog odgovora na osnovu unetog ulaznog teksta korisnika. Jedan od načina za rešavanje problema predikcije sekvenci je korišćenje rekurentne neuronske mreže.

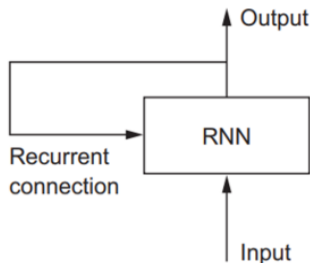
A. Rekurentna neuronska mreža

RNN (*Recurrent Neural Network*) je vrsta neuronskih mreža gde se izlaz neurona vraća na njegov ulaz i to kreira

⁵ <https://www.kaggle.com/datasets/tusharkhete/dataset-for-medicalrelated-chatbots>

⁶ <https://spacy.io/>

neku vrstu internog stanja mreže, odnosno memoriju. Rekurentnost znači da će izlaz trenutne iteracije biti ulaz za narednu iteraciju. Na slici 1 ilustrovana je struktura RNN mreže. Za svaki element iz sekvence, posmatra se trenutni ulaz i prethodno stanje. Ipak klasična RNN nije idealno rešenje. Teška je za obučavanje zbog problema nestajućeg gradijenta (eng. *vanishing gradient*), gde se sa povećavanjem podataka za obučavanje, mreža sve manje obučava. LSTM (*Long-Short Term Memory*) je varijanta RNN mreže koja omogućava rešenje ovog problema, a takođe ima bolju preciznost od klasične RNN [16].



Slika 1. Struktura rekurentne neuronske mreže

B. Long short-term memory

LSTM je najčešće korišćen model neuronske mreže za implementaciju *chatbot-a*. Najbitnija prednost LSTM je mogućnost dugoročnog pamćenja sekvenci. Dugoročno pamćenje sekvenci je omogućeno preuzimanjem i pamćenjem konteksta dela sekvence tokom čitave obrade ulazne sekvence. LSTM rešava problem nestajućeg gradijenta korišćenjem nekoliko kapija kako bi se kontrolisao protok informacija koje ulaze i izlaze iz memorijske ćelije. LSTM se sastoji iz internog stanja ćelije (eng. *cell state*), koje predstavlja dugoročnu memoriju i 3 kapije koje kontrolišu količinu informacija koje prolaze kroz stanje ćelije. Postoje ulazna kapija (eng. *input gate*), kapija za zaboravljanje informacija (eng. *forget gate*) i izlazna kapija (eng. *output gate*). Ulazna kapija utiče na odabir informacija koje se dodaju u stanje ćelije, kapija za zaboravljanje informacija kontrolišu koje informacije treba da se zadrže u stanju ćelije, a koje da se zaborave, a izlazna kapija kontrolišu koje informacije treba proslediti na izlaz stanja ćelije. Svaka kapija predstavlja sigmoidalnu funkciju koja vraća vrednosti između 0 i 1, ukazujući na to koliko informacija treba propustiti kroz kapiju.

C. Seq2seq model, enkoder-dekoder arhitektura

Za implementaciju našeg *chatbot-a*, koristi smo seq2seq model. Ovo je model čija se arhitektura sastoji iz dve komponente - enkodera i dekodera. Između enkoder i dekoder sloja, implementirali smo mehanizam pažnje, kako bi se na što efikasniji način uvideo kontekst ulazne rečenice.

Pri implementaciji enkodera, za generisanje reprezentacije ulazne rečenice koristili smo bidirekcionu LSTM. Bidirekcionu LSTM se sastoji iz dve LSTM mreže, jedne za prolaz unapred (eng. *forward pass*) koja čita ulaznu sekvencu s leva na desno i druge mreže za prolaz unazad (eng. *backward pass*) koja čita sekvencu s desna na levo. Ovo omogućava enkoderu da na najefikasniji način shvati kontekst ulazne rečenice. Bidirekcionu LSTM je idealno rešenje za dugačke sekvence.

Pored bidirekcionu LSTM mreže, enkoder poseduje *embedding* sloj koji pretvara reči u vektore fiksne dužine.

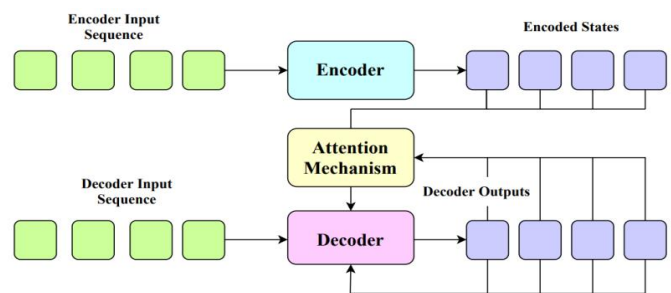
Enkoder funkcioniše tako što uzima čitavu rečenicu i procesira je reč po reč, odnosno sekvencu po sekvencu. Pri obradi svake pojedinačne sekvence, bidirekcionu LSTM vraća skriveno stanje (eng. *hidden state*) i stanje ćelije (eng. *cell state*). Svako skriveno stanje trenutne sekvence utiče na skriveno stanje sledeće sekvence, a krajnje skriveno stanje se posmatra kao rezime cele rečenice. Ovo poslednje stanje se zove kontekst vektor (eng. *thought vector*) i on predstavlja smisao, odnosno rezime cele rečenice.

Dekoder kao ulaz koristi kontekst vektor generisan od strane enkodera. Pored toga, dekoder koristi LSTM mrežu i vrši predikciju izlazne sekvence.

D. Mehanizam pažnje

Enkoder-dekoder arhitektura funkcioniše dobro kada su u pitanju kraće rečenice, odnosno kraći nizovi sekvenci. U slučaju dužih rečenica, postoji nemogućnost ekstrakcije konteksta sekvence. Zbog toga se koristi mehanizam pažnje, koji je dobio naziv zbog mogućnosti pridavanja značaja relevantnim sekvencama. Mehanizam pažnje omogućava dekoderu da prilikom predikcije izlazne sekvence selektivno obraća pažnju na ulazne sekvence. Kod klasične enkoder-dekoder arhitekture, za svaku sekvencu koju učitava, enkoder vrati *intermediate output*, ali taj izlaz se ne čuva dugoročno nego se koristi samo skriveno stanje i stanje ćelije enkodera za obradu sledeće sekvence. Prilikom korišćenja mehanizma pažnje, čuvaju se *intermediate output*-i iz enkodera koji sa sobom nose određeni nivo značajnosti i konteksta. Model se trenira tako da uči da obraća selektivnu pažnju na *intermediate output*-e i da uviđa povezanosti između njih i izlaznih sekvenci.

U radu je korišćen *Bahdanau* mehanizam pažnje, jer on dobro funkcioniše sa dugačkim sekvencama. Vršu linearnu kombinaciju stanja enkodera i dekodera. Na slici 2, prikazana je struktura enkoder-dekoder arhitekture sa mehanizmom pažnje.



Slika 2. Struktura enkoder-dekoder arhitekture sa mehanizmom pažnje

E. Adam – algoritam za optimizaciju

Za optimizaciju performansi, koristili smo Adam (*Adaptive Moment Estimation*) Optimizator [17]. Ovaj algoritam optimizacije koristi najefikasnije delove od *Momentum* i *RMSProp* (*Root Mean Squared Propagation*) algoritama.

Adam dobija brzinu od *Momentum* algoritma i mogućnost da prilagodi gradijente u različitim pravcima od RMSProp algoritma. Ova kombinacija čini Adam algoritam vrlo moćnim a stoga i često korišćenim algoritmom optimizacije u mašinskom učenju.

F. Inkrementalno obučavanje

Jedan od izazova koji se javljaju u primeni navedenog modela je problem nedostatka memorije (eng. *memory constraint problem*), koji se javlja usled veličine ulaznih podataka. Jedan od načina za rešavanje ovog problema je inkrementalno obučavanje modela. Inkrementalno obučavanje predstavlja metodu koja omogućava neprekidno učenje modela, što dovodi do proširenja njegovog znanja i poboljšanja performansi. Međutim, ovaj pristup sa sobom nosi rizik gubitka prethodno stečenog znanja pri obučavanju modela novim podacima (eng. *catastrophic forgetting problem*). Za rešavanje ovog problema predlažemo pristup koji se sastoji iz tri faze. Prva faza obuhvata obučavanje modela na originalnom skupu podataka kako bismo kreirali početni model. U drugoj fazi, pripremamo model za inkrementalno obučavanje tako što zamrzavamo težine već kreiranog modela i dodajemo novi sloj (npr. *Dense layer*) za obradu novih podataka. Nakon obučavanja modela na novim podacima, prelazimo na treću fazu koja obuhvata *fine-tuning* proces, pri kojem model dodatno obučavamo kombinacijom starih i novih podataka kako bi se osiguralo da su težine prilagođene za oba skupa podataka. Cilj ovog pristupa je obezbeđivanje ravnoteže između proširenja znanja i očuvanja prethodno stečenog znanja. Implementacija ovog pristupa može biti izazovna, ali omogućava efikasno i pouzdano inkrementalno obučavanje modela, bez rizika od gubitka prethodno stečenog znanja.

IV. REZULTATI I DISKUSIJA

Namena *chatbot*-a je simulacija konverzacije između pacijenta i lekara, u cilju predikcije dijagnoze pacijenta ili davanja preporuke za lečenje pacijenata. Skupovi podataka emuliraju strukturu realne konverzacije između lekara i pacijenta, gde pacijent najpre postavlja pitanje, odnosno navodi svoj problem, a lekar zatim daje odgovor. Za obučavanje modela, korišteni su skupovi podataka opisani u 2. poglavlju. Skupovi podataka su u json formatu, gde se svaki json objekat sastoji iz pitanja pacijenta i odgovora lekara, koji su po potrebi sumariзовани i samim tim adekvatne dužine za ulaz u enkoder-dekoder model.

U okviru ovog poglavlja, opisaćemo i upoređićemo nekoliko konfiguracija *chatbot*-a. Kao što smo već spomenuli u 2. poglavlju, vršićemo nezavisno obučavanje *chatbot*-a, sa dva različita skupa podataka (*MedDialogPlus* i *Intents*), koji ne poseduju zajedničke podatke. U skupu podataka *MedDialogPlus*, odgovori lekara su namenjeni isključivo za navedeno pitanje pacijenta. Odgovori ne predstavljaju sveobuhvatno rešenje za datu tematiku jer nisu uopšteni. S druge strane, *Intents* skup podataka je takav da sa slična pitanja daje uvek potpuno isti odgovor. U ovom skupu podataka, odgovori su generalizovani, uopšteni i tiču se bilo kakvog problema iz iste oblasti.

A. BLEU metrika

Za evaluaciju performansi, koristili smo BLEU metriku, koja je u skorije vreme postala tipična metrika za evaluaciju performansi *chatbot*-a. Vrednost BLEU metrike zavisi od kvaliteta trening podataka, kompleksnosti modela i kompleksnosti zadatka za koji je *chatbot* kreiran da obavlja. BLEU metrika vraća vrednosti između 0 i 1, gde veće vrednosti ukazuju na veću sličnost između generisanog i očekivanog teksta. BLEU vrednost je bazirana na podudaranju n-grama. N-grami predstavljaju sekvence reči, gde je n broj reči sekvence. Iako je BLEU metrika popularna za evaluaciju performansi *chatbot*-a, vrlo je izazovno evaluirati ovakav sistem zbog same prirode *chatbot*-a. Bitno je uzeti u obzir ljudsku evaluaciju kako bi se što bolje procenio kvalitet *chatbot*-a. Da bismo koristili BLEU metriku, potrebno je da ulazne podatke podelimo na trening i test skup.

Za oba spomenuta skupa podataka (*MedDialogPlus* i *Intents*), izvršena je podela na trening i test skup, gde je veličina test skupa 30%. Za veličinu test skupa obično se uzimaju vrednosti između 20% i 30%, zavisno od veličine obučavajućeg skupa podataka ali i problema koji se rešava. Za veće skupove podataka, za veličinu test skupa obično se uzima vrednost od 30%, dok se za manje skupove podataka obično uzimaju manje vrednosti. Podela na trening i test skup je neohodna jer želimo da odvojimo podatke nad kojima se vrši obučavanje modela, od podataka nad kojim se vrši testiranje modela. Na taj način, uviđamo realne rezultate performansi *chatbot*-a. Sve tehnike pretprocesiranja podataka koje su primenjene nad skupom podataka za treniranje modela, primenjene su i nad test skupom podataka.

B. Tačnost i funkcija cilja

Pored BLEU metrike, za evaluaciju performansi, posmatrali smo tačnost (eng. *accuracy*) i funkciju cilja (eng. *loss function*). Tačnost je metrika koja se odnosi na procenat tačnih predikcija koje model napravi nad zadatim skupom podataka. Visoka tačnost ukazuje da model vrši tačna predviđanja. Funkcija cilja je metrika koja predstavlja razliku između predikcije modela i stvarne vrednosti. Za računanje funkcije cilja u našem projektu, koristili smo *Categorical Entropy* funkciju. Niska vrednost funkcije cilja ukazuje da se model dobro obučava nad skupom podataka.

C. Pregled postojećih konfiguracija

U nastavku ćemo komentarisati sledeće rezultate:

- konfiguracije *chatbot*-a obučenog nad *MedDialogPlus* skupom podataka, bez tehnika inkrementalnog obučavanja;
- konfiguracije *chatbot*-a obučenog nad *MedDialogPlus* skupom podataka, koristeći inkrementalno obučavanje;
- konfiguracije *chatbot*-a obučenog nad *Intents* skupom podataka, bez tehnika inkrementalnog obučavanja;
- konfiguracije *chatbot*-a obučenog nad *Intents* skupom podataka, koristeći inkrementalno obučavanje.

D. Pregled konfiguracija koje ne koriste tehnike inkrementalnog obučavanja

U tabeli 1, prikazano je nekoliko konfiguracija modela *chatbot*-a, koji je usled problema nedostatka memorije, obučen samo nad delom *MedDialogPlus* skupa podataka.

	config 1	config 2	config 3
broj epoha	100	100	30
batch size	24	24	24
Mean bleu (1-gram)	0.139684	0.138339	0.141693
Mean bleu (4-gram)	0.00202	0.002057	0.002394

Tabela 1. Nekoliko konfiguracija *chatbot*-a, obučenog nad delom *MedDialogPlus* skupa podataka

U tabeli 2 nalazi se nekoliko konfiguracija *chatbot*-a, koji je obučen nad *Intents* skupom podataka.

	config 1	config 2	config 3	config 4	config 5
broj epoha	100	70	70	100	110
batch size	24	24	64	64	64
Mean bleu (1-gram)	0.6775	0.6366	0.3469	0.6270	0.6028
Mean bleu (4-gram)	0.6158	0.5673	0.1696	0.5329	0.5135

Tabela 2. Nekoliko konfiguracija *chatbot*-a, obučenog nad *Intents* skupom podataka

Posmatrajući BLEU vrednosti iz tabela 1 i 2, uvidamo znatnu razliku u performansama *chatbot*-a obučenog nad *MedDialogPlus* skupom podataka i *chatbot*-a obučenog nad *Intents* skupom podataka. Prosečna BLEU vrednost 1-grama najbolje konfiguracije *chatbot*-a obučenog nad *MedDialogPlus* skupom podataka iznosi 0.141693, dok je najbolja prosečna vrednost 1-grama za *Intents* skup podataka čak 0.677456.

Razlog za nižu vrednost BLEU metrike *chatbot*-a obučenog nad *MedDialogPlus* skupom podataka je to što su pitanja pacijenata vrlo individualizovana. Svaki pacijent formuliše pitanje u potpuno slobodnoj formi. Pored svojih simptoma, neki pacijenti pišu i o svom životu, godinama, porodičnoj istoriji, prpratnim životnim problemima ili čak tegobama na poslu. Sumarizacijom pitanja na odgovarajuću dužinu, može se izgubiti suština pitanja. Sumarizacijom pitanja svodimo sve slične probleme na slično ili isto pitanje, ali svaki odgovor lekara, i pored sumarizacije odgovora, je individualno namenjen inicijalnom pitanju pacijenta. U *MedDialogPlus* skupu podataka, lekar u svom odgovoru pokazuje empatiju prema pacijentu, pruža nadu i utehu, ali zbog toga, vrlo je zahtevno izvršiti sumarizaciju u kojoj ostane samo suština. Mreži se na taj način, za slična ili ista pitanja, dovode potpuno različiti odgovori, a to doprinosi nekozistentnosti pri obučavanju. Ovo predstavlja glavni

nedostatak *MedDialogPlus* skupa podataka i razlog za nižu vrednost BLEU metrike.

S druge strane, *Intents* skup podataka poseduje odgovore uopštene za bilo koje pitanje iz slične oblasti. Za ista ili slična pitanja, na mrežu se uvek dovodi isti odgovor. Zbog toga, konfiguracije *chatbot*-a koje za obučavanje koriste *Intents* skup podataka, daju znatno bolje rezultate.

E. Pregled konfiguracija koje koriste tehnike inkrementalnog obučavanja

U tabeli 3 prikazane su dve konfiguracije, u kojima se koristi tehnika inkrementalnog obučavanja. Prva konfiguracija opisuje inkrementalno obučavanje nad *MedDialogPlus* skupom podataka, dok druga konfiguracija opisuje inkrementalno obučavanje korišćenjem *Intents* skupa podataka. Parametri su navedeni s obzirom na faze obučavanja, zajedno sa brojem podataka koji su korišteni u tim fazama. Tokom inkrementalnog obučavanja, model se u prvoj fazi obučava nad većim delom skupa podataka, pri čemu je arhitektura modela ista kao u prethodno navedenim konfiguracijama. U drugoj fazi, model se doobučava na preostalim podacima, pri čemu se dodaje još jedan sloj, u ovom slučaju *Dense* sloj, na prethodno opisanu arhitekturu. Broj epoha i veličina paketa (eng. *batch size*) su usklađeni tokom *fine-tuning* procesa kako bi se postigao balans između već stečenog znanja i novih informacija.

	config 1		config 2	
	I Faza	II Faza	I Faza	II Faza
broj podataka	4050	1002	135	15
broj epoha	100	85	100	70
batch size	24	24	24	18
Mean bleu (1-gram)	0.138483		0.790685	
Mean bleu (4-gram)	0.001295		0.743953	

Tabela 3. Konfiguracije *chatbot*-a, u kojima se koristi inkrementalno obučavanje

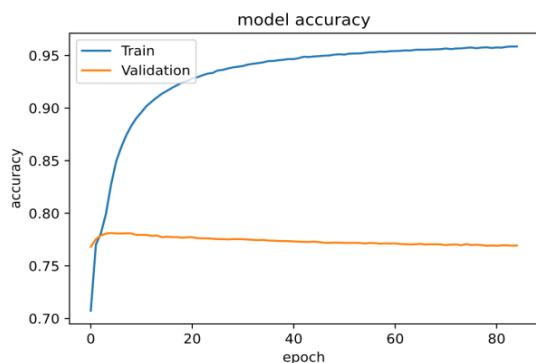
F. Diskusija rezultata

Razmatrajući vrednosti BLEU metrike iz tabele 3, primetno je da se vrednosti za modele obučene nad *MedDialogPlus* skupom podataka ne razlikuju značajno kada se koristi inkrementalno obučavanje i kada se ne koristi inkrementalno obučavanje. Najbolja prosečna vrednost BLEU metrike za 1-gram modela obučenog nad *MedDialogPlus* skupom bez korišćenja inkrementalnog obučavanja je 0,141, dok je najbolja prosečna vrednost BLEU metrike za 1-gram modela obučenog nad *MedDialogPlus* skupom podataka sa korišćenim tehnikama inkrementalnog obučavanja 0,138. Iako je rezultat nešto niži kod inkrementalnog obučavanja, možemo zaključiti da je proces inkrementalnog obučavanja uspešno sproveden i izbegnut je gubitak već stečenog znanja. Mogući napredak u rezultatu inkrementalnog obučavanja ostvariv je dužim procesom *fine-tuning*-a, kao i uvođenjem treće faze koja predstavlja obučavanje nad starim i novim podacima.

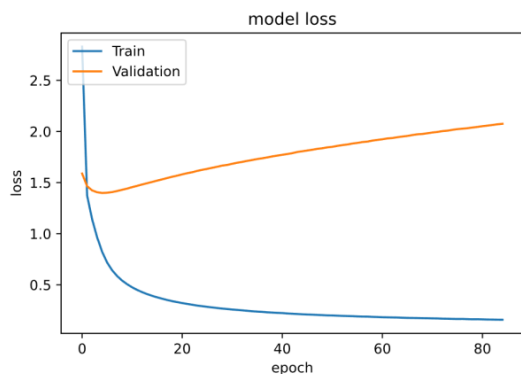
Ako uporedimo rezultate za model koji je obučen nad *Intents* skupom podataka (*config* 2 iz [tabele 3](#)) korišćenjem tehnike inkrementalnog obučavanja sa modelom obučenim nad istim skupom podataka bez korišćenja tehnike inkrementalnog obučavanja (konfiguracije iz [tabele 2](#)), primećujemo napredak. Bolje rezultate uviđamo za vrednosti BLEU metrike za 4-gram, gde je vrednost bez inkrementalnog obučavanja 0.615758 ([tabela 2](#)), dok je vrednost BLEU metrike sa inkrementalnim obučavanjem 0.743953. Napredak u rezultatu može se pripisati većem broju epoha i složenijoj krajnjoj arhitekturi, ali i konciznim ulaznim podacima.

Na slikama [3](#) i [4](#) nalaze se grafici za prikaz tačnosti i funkcije cilja za konfiguraciju *config* 1 iz [tabele 3](#), koja predstavlja inkrementalno obučeni model nad *MedDialogPlus* skupom podataka. Na slikama [5](#) i [6](#) prikazani su tačnost i funkciju cilja za konfiguraciju *config* 2 iz [tabele 3](#), koja predstavlja inkrementalno obučeni model nad *Intents* skupom podataka.

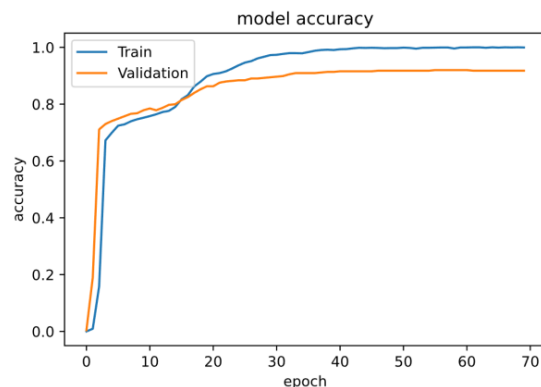
Na slikama [3](#) i [5](#) vidimo trend rasta tačnosti za trening skup, dok na slikama [4](#) i [6](#) uviđamo trend opadanja funkcije cilja za trening skup. Za model koji je inkrementalno obučen nad *Intents* skupom podataka, vidimo da tačnost raste i za validacioni skup podataka, a funkcija cilja opada (slike [5](#) i [6](#)). Ipak, kod modela koji je inkrementalno obučen nad *MedDialogPlus* skupom podataka, sa [slikom 3](#) i [4](#), vidimo da to nije slučaj. Dovođenje različitih odgovora za vrlo slična ili ista pitanja, doprinosi nekozistentnom obučavanju neuronske mreže. Usled korišćenja neusaglašenih podataka za obučavanje mreže, dolazi do pada performansi *chatbot*-a.



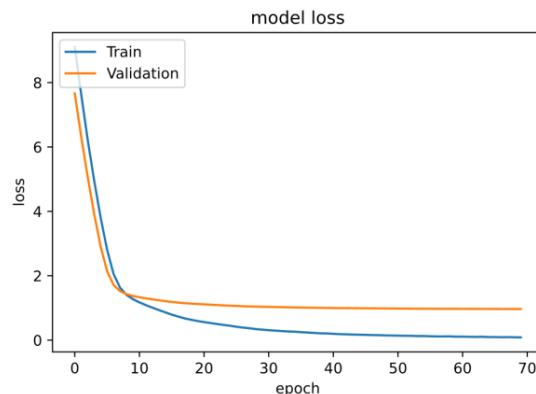
Slika 3. Tačnost za konfiguraciju *config* 1 iz tabele 3



Slika 4. Funkcija cilja za konfiguraciju *config* 1 iz tabele 3



Slika 5. Tačnost za konfiguraciju *config* 2 iz tabele 3



Slika 6. Funkcija cilja za konfiguraciju *config* 2 iz tabele 3

G. Poređenje BLEU metrike sa sličnim radovima

Analiziranjem BLEU vrednosti sličnih radova, vidimo da su autori rada [\[7\]](#) za model u kojem je korištena seq2seq arhitektura i mehanizam pažnje dobili BLEU vrednost 1-grama 0.137, što je slično našoj vrednosti inkrementalno obučenog modela nad *MedDialogPlus* skupom podataka. Rad [\[9\]](#) ima nešto bolju vrednost BLEU metrike, koja iznosi 0.2529. BLEU metrika 1-grama dobijena u radu [\[18\]](#) iznosi čak 0.7, što je vrlo sličan rezultat našoj vrednosti BLEU metrike za model obučen nad *Intents* skupom podataka, bez tehnika inkrementalnog obučavanja. Uzimajući u obzir veliki opseg vrednosti rezultata BLEU metrike, dolazimo do zaključka da metrika najviše zavisi od kvaliteta ulaznih podataka.

H. Ograničenja

Najveće ograničenje prilikom implementacije *chatbot*-a je skup podataka koji se koristi za obučavanje. Nepostojanje adekvatnog skupa podataka umanjuje efikasnost performansi *chatbot*-a, što dovodi do nemogućnosti obezbeđivanja tačnih preporuka i dijagnoza. Takođe, nedostatak standardizacije u terminologiji različitih sajtova odakle su podaci preuzimani, otežava integraciju podataka sa različitih izvora. Postojanje generalizovanog i konciznog skupa podataka, koji je fokusiran na suštinu pitanja i odgovora, znatno poboljšava performanse *chatbot*-a, što vidimo na osnovu rezultata *chatbot*-a obučenog nad *Intents* skupom podataka.

Bitno je naglasiti prirodu podataka s kojim naš *chatbot* radi. Većina radova koji su bazirani na istoj arhitekturi kao i naš *chatbot*, rade sa sekvencama dužine između 5 i 15 reči, dok naš *chatbot* radi sa sekvencama dužine između 5 i 160 reči. Dok su u drugim sličnim radovima izbačene rečenice koje su duže od 15 reči, mi smo radili sumarizaciju teksta i ograničili smo maksimalnu dužinu teksta na 160 reči. Stoga, naš *chatbot* je obučen da radi sa dužim sekvencama i zato ima mogućnost generisanja boljeg odgovora. Za rukovanje sa dužim sekvencama, ali i velikom količinom podataka, neophodna je veća količina RAM (*Random-Access Memory*) memorije, zbog čega smo suočeni sa problemom nedostatka memorije. Ovaj problem smo u određenoj meri regulisali primenom inkrementalnog obučavanja.

V. ZAKLJUČAK

U radu je predložena arhitektura *chatbot* sistema koji je razvijen s ciljem pružanja pomoći pacijentima pre posete lekara. Kreirani sistem bi imao veliki doprinos u zdravstvenom sistemu, jer bi bio dostupan 24 sata, 7 dana u nedelji, što bi omogućilo pacijentima brz ali pre svega siguran izvor informacija i saveta u vezi sa njihovim zdravstvenim problemima. Ovo bi moglo poboljšati pristup zdravstvenoj nezi, smanjiti opterećenost medicinskog osoblja i omogućiti bržu i efikasniju komunikaciju sa pacijentima, posebno u hitnim situacijama. Ipak, ovakav sistem nema kao cilj zamenu medicinskog osoblja i stručnih dijagnoza, nego predstavlja alat za povećanje njihove efikasnosti. Za ozbiljne stručne dijagnoze potrebno je konsultovati stručnjaka iz određene oblasti, zavisno od vrste problema.

Navedeni sistem je implementiran pomoću seq2seq modela koji se sastoji iz enkoder i dekode sloja. Pri implementaciji enkodera korišten je bidirekcionni LSTM, dok se za implementaciju dekode sloja koristi LSTM. Između enkoder i dekode sloja implementiran je *Bahdanau* mehanizam pažnje. Prednost opisane arhitekture *chatbot*-a, ogleda se u mogućnosti rukovanja dugim sekvencama ulaznih podataka čime se povećava mogućnost davanja tačne i detaljne dijagnoze ili preporuke za lečenje. Dodatno, ovaj rad pruža uvid u mogućnost prevazilaženja problema nedostatka memorije implementacijom inkrementalnog obučavanja. Ostvareni rezultati u radu, u kojima se primeti neznatna razlika vrednosti BLEU metrike između modela obučenog pomoću tehnika inkrementalnog obučavanja i modela obučenog na tradicionalni način, dokazuju da je moguće uspešno prevazići problem gubitka prethodno stečenog znanja, koji predstavlja najveću prepreku prilikom korištenja inkrementalnog obučavanja. Međutim, glavni problem pri implementaciji ovog sistema je nedostatak dovoljno obimnog i konciznog skupa podataka. Kako bi *chatbot* mogao da daje tačne dijagnoze i preporuke za lečenje, potrebno je obučiti sistem na još većem skupu podataka. Iako je obim skupa podataka izuzetno važan, najbitniji je kvalitet i priroda skupa podataka. Pitanja i odgovori treba da budu što koncizniji, precizniji i da su fokusirani na suštinu problema. Kreiranje takvog skupa podataka, koji je adekvatnog obima, znatno bi popravilo performanse sistema.

Jedan pravac za budući razvoj *chatbot*-a je mogućnost pamćenja konverzacije sa pacijentom i generisanje odgovora

na osnovu celokupne posmatrane istorije razgovora. Pamćenjem konverzacije bila bi omogućena kvalitetnija dijagnoza jer se sagleda celokupna slika problema pacijenta. Dodatno, za dalji pravac razvoja, preporučujemo upotrebu učenja sa potkrepljenjem kako bi *chatbot* ispitivao pacijenta o implicitnim simptomima i na taj način dobio jasniju sliku o navedenom problemu.

Reference

- [1] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine", doi: <https://doi.org/10.1145/357980.357991>.
- [2] E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: <https://doi.org/10.1038/323533a0>.
- [3] Sojasingarayar, "Seq2Seq AI Chatbot with Attention Mechanism," *Artif. Intell.*, 2020, doi: <https://doi.org/10.48550/arXiv.2006.02767>.
- [4] Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." *arXiv*, Jun. 15, 2011. Accessed: May 07, 2023. [Online]. Available <http://arxiv.org/abs/1106.3077>
- [5] Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks".
- [6] Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate." *arXiv*, May 19, 2016. Accessed: May 07, 2023. [Online]. Available: <http://arxiv.org/abs/1409.0473>
- [7] R. Goel, S. Vashisht, A. Dhanda, and S. Susan, "An Empathetic Conversational Agent with Attentional Mechanism," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1–4. doi: [10.1109/ICCCI50826.2021.9402337](https://doi.org/10.1109/ICCCI50826.2021.9402337).
- [8] P. F. Brown *et al.*, "A STATISTICAL APPROACH TO MACHINE TRANSLATION," vol. 16, no. 2, 1990.
- [9] N. A. I. Omeregbe, I. O. Ndamani, S. Misra, O. O. Abayomi-Alli, and R. Damaševičius, "Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic," *J. Healthc. Eng.*, vol. 2020, pp. 1–14, Sep. 2020, doi: <https://doi.org/10.1155/2020/8839524>.
- [10] N. Chomsky, "Three models for the description of language," *IRE Trans. Inf. Theory*, vol. 2, no. 3, pp. 113–124, 1956, doi: [10.1109/TIT.1956.1056813](https://doi.org/10.1109/TIT.1956.1056813).
- [11] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 464–471, 2002, doi: <https://doi.org/10.1109/72.991432>.
- [12] Z. Wei *et al.*, "Task-Oriented Dialogue System for Automatic Diagnosis", doi: <https://doi.org/10.48550/arXiv.2004.14254>.
- [13] L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, 1959, doi: <https://doi.org/10.1147/rd.33.0210>.
- [14] X. He *et al.*, "MedDialog: Two Large-scale Medical Dialogue Datasets." *arXiv*, Jul. 07, 2020. Accessed: May 07, 2023. [Online]. Available: <http://arxiv.org/abs/2004.03329>
- [15] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. doi: <https://doi.org/10.3115/v1/D14-1162>.
- [16] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [17] P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." *arXiv*, Jan. 29, 2017. Accessed: May 07, 2023. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [18] B. Rakib, E. A. Rumky, A. J. Ashraf, Md. M. Hillas, and M. A. Rahman, "Mental Healthcare Chatbot Using Sequence-to-Sequence Learning and BiLSTM," in *Brain Informatics*, M. Mahmud, M. S. Kaiser, S. Vassanelli, Q. Dai, and N. Zhong, Eds., Cham: Springer International Publishing, 2021, pp. 378–387, doi: https://doi.org/10.1007/978-3-030-86993-9_34