Predlog projekta iz predmeta
„Sistemi za istraživanje i analizu podataka“
Healthcare chatbot

1. Jasna definicija cilja projekta

Cilj projekta je razvoj conversational chatbot-a za predikciju dijagnoze ili davanje preporuka za lečenje pacijenata na osnovu simptoma koje pacijent navede. Conversational chatbot je vrsta chatbota 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.

2. Motivacija problema rešavanog u projektu

Opšte zdravsteno 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 osnovu dijagnozu pacijenta i da mu da preporuke pre konsultacije sa lekarom. Na osnovu članka CNBC sajta (https://www.cnbc.com/2022/06/22/100-million-adults-have-health-care-debt-and-some-owe-10000-or-more.html), preko 100 miliona Amerikanaca ima dugove koji se tiču zdravstene zaštite, a chatbot bi imao kao cilj da pomogne u smanjenju troškova zdravstvene zaštite ali bi istovremeno omogućio i veću pristupačnost medicinskom znanju.

3. Relevantna literatura

[1] Abonia Sojasingarayar, Seq2Seq AI Chatbot with Attention Mechanism

1. Cilj - Razvoj conversational chatbot-a koji generiše odgovor na osnovu unesenih ulaznih podataka, u cilju simulacije ljudskog razgovora.
2. Metodologija - Recurrent Neural Network (RNN), Sequence-To-Sequence model (Seq2Seq se sastoji iz encoder-a i decoder-a koji predstavljaju LSTM tip rekurentne neuronske mreže)
3. Skup podataka - Korišćen Cornell Movie Dialog Corpus Dataset, koji sadrži kolekciju razgovora izvučenih iz scenarija filmova. Skup podataka sadrži 220,579 razgovora između 9,035 likova iz 617 filmova. https://www.cs.cornell.edu/~christian/Cornell_MovieDialogs_Corpus.html
4. Softver - Korišten tensorflow V1.14.0 za razvoj Seq2Seq modela.
5. Evaluacija - Izvršeno poređenje 3 modela, odnosno 3 konfiguracije sa različitom kombinacijom hiperparametara. Za određivanje najboljeg modela korištena je ljudska estimacija.
6. Rezultat – Najpre je izvršeno pretpocesiranje skupa podataka pri čemu su izbačeni metapodaci (movie ID, character ID…), separatori, omogučena je podrska za UTF-8 standard i podaci su generalno prečišćeni. Nakon toga, podaci su podeljeni u dve liste kako bi bio zadovoljen format ulaza za Seq2Seq model. Prva lista predstavlja pitanja, a druga odgovore.
   Zatim su formirane tri kongifuracije Seq2Seq modela sa različitim kombinacijama hiperparametara. Svaka od konfiguracija je dala različite rezultate, ali je 3. kofiguracija nadmašila druge dve. Ona je imala najmanji broj epoha (50, dok su prva i druga konfiguracija imale 500 i 100 epoha), najmanji batch size (32, dok su

prva i druga konfiguracija imale batch size 128 i 512) i najveću vrednost za parametar rnn size (1024, dok su prva i druga konfiguracija imale rnn size 128 i 512). Performanse se mogu dodatno poboljšati odabirom drugih vrednosti za hiperparametre.

7. Zaključak - Sama ideja kreiranja conversational chatbota će biti predmet rada u našem projektu. Naš projekat će koristiti metodologiju opisanu u ovom naučnom radu - RNN mrežu, Sequence To Sequence model (encoder-decoder implementirani kao LSTM). Što se tiče evaluacije, jedan od načina evaluacije našeg projekta biće ljudska estimacija, koja je takođe korištena u ovom naučnom radu.

8. Nedostaci – Poboljšanje skupa podataka bi dovelo do boljih rezultata.

[2] Qianlong Liu, Zhongyu Wei, Baolin Peng, Xiangying Dai, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, Task-oriented Dialogue System for Automatic Diagnosis

1. Cilj - Kreiranje dijalog sistema (chatbot-a) za automatsku dijagnozu pacijenata na osnovu unesenih simptoma, analizom prethodno sakupljenih podataka. Prilikom konverzacije sa pacijentom, sistem je osposobljen da postavlja pitanja o propratnim simptomima kako bi bila izvršena tačna dijagnoza (korišćenjem reinforcement learning-a).

2. Metodologija – NLP, NLU, NLG, Markov Decision Process (MDP), Deep Q-network (DQN), Reinforcement Learning (RL)

3. Skup podataka - Autori projekta su napravili svoj dataset koji se sastoji od dve vrste podataka. Skup podataka prikupljen je sa kineskih medicinskih online foruma. Prvi skup podataka su podaci koje pacijent u toku pregleda 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.

4. Evaluacija – Izvršeno je poređenje performansi DQN Agenta, Rule-based Agenta i Random Agenta. Za poređenje ovih modela korišćeni su success rate (stopa uspeha), average reward (prosečna nagrada, specifična mera korišćena kod MDP) i average number of turns per dialogue session (prosečan broj puta koliko agent i pacijent razmene poruke, na nivou jedne konverzacije).
   Izvršeno je i poređenje klasifikacionog modela SVM samo sa ekplicitnim podacima i klasifikacionog modela SVM sa eksplicitnim i implicitnim podacima kroz meru accuracy (tačnost).

5. Rezultat – DQN Agent se pokazao kao najprecizniji, sa 65% success rate, zatim Rule-based Agent sa 23% i na kraju Random Agent sa 6% success rate.
   Sto se tiče tačnosti SVM modela, rezultati pokazuju da postojanje implicitnih simptoma povećava tačnost pri dijagnozi bolesti za više od 10%. To govori da se korišćenjem reinforcement learning-a moze izvršiti preciznija dijagnostika bolesti.

6. Zaključak – Sama ideja kreiranja conversational chatbota će biti predmet rada u našem projektu.
   Proces obrade podataka NLP tehnikama će biti iskorišten i u našem projektu. S obzirom da je tematika ovog naučnog rada i našeg projekta ista, korisno je odraditi dublju analizu načina na koji je korišćen skup podataka iz ovog naučnog rada. Korišćenje implicitih podataka, pored eksplictnih podataka, navelo nas je na šire istraživanje o dataset-u koji ćemo koristiti u projektu, tako da dataset sadrži i konverzacije u kojima doktor ispituje pacijente o njihovim dodatnim simptomima (implicitno prikupljanje simptoma).
   Korišćenje reinforcement learning-a za poboljšanje performansi, ostaje kao ideja za buduću nadogradnju našeg chatbota.

7. Nedostaci - Veličina dataseta je oko 5000 dijaloga, što je nedovoljno za adekvatno obučavanje DQN-a. Zato je success rate DQN-a oko 65%.

[3] Nicholas A. I. Omoregbe, Israel O. Ndaman, Sanjay Misra, Olusola O. Abayomi-Alli and Robertas Damaševičius, Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic

1. Cilj - Korišćenje NLP tehnika i fuzzy SVM (Support Vector Machine) klasifikatora za razvoj conversational chatbot-a za predviđanje dijagnoze pacijenta na osnovu unetih simptoma.
2. Metodologija – NLP tehnike, fuzzy SVM
3. Skup podataka – Korišćena su 4 izvora podataka.
    1. WordNet – leksička baza podataka semantičkih odnosa između reci (https://wordnet.princeton.edu/)
    2. YAGO (Yet Another Great Ontology) – baza podataka korišćena za konstrukciju 'knowledge graph-a' (https://yago-knowledge.org/)
    3. UMLS (Unified Medical Language System) – baza medicinskih izraza, za međusobno povezivanje medicinskih termina (https://www.nlm.nih.gov/research/umls/index.html)
    4. Disease Ontology (DO) - baza podataka sa preko 10,000 bolesti (https://disease-ontology.org/)
4. Evaluacija – Za evaluaciju performansi sistema korišćena je BLEU (Bilingual Evaluation Understudy) metrika. Za određivanje upotrebljivosti sistema (SUS – System Usability Scale), pozvan je određeni broj pacijenata i zdravstvenih radnika da testiraju sistem i ostave svoj utisak.
5. Rezultat – Napre se vrši procesiranje teksta NLP tehnikama - noise removal (uklanjanje nerelevantnih podataka), tokenization, parsing (POS – Part of speech). Zatim se vrši obučavanje SVM klasifikatora na osnovu prethodno procesiranih podataka NLP tehnikama.
   Rezultat SUS evaluacije sistema je 80.4%, a minimalni prag za ocenu 'odličan' na SUS skali je 80.3%.
6. Zaključak – NLP tehnike spomenute u ovom naučnom radu, biće iskorištene i u okviru našeg projekta za razumevanje teksta i razmatranje adekvatnog odgovora. Za evaluaciju performansi našeg projekta biće korištena BLEU metrika, a za određivanje nivoa upotrebljivosti sistema koristićemo SUS metriku namenjenu za chatbot-e – CUQ (Chatbot Usability Questionary).

4. Skup podataka

Da bismo naš chatbot obučili sa što većom količinom podataka, odlučili smo da kombinujemo podatake sa više izvora. Nad pronađenim podacima vršiće se dalja analiza kako bismo uklonili nepotrebne delove i redudantne podatake.

Do jednog skupa podataka došli smo nakon čitanja naučnog rada https://arxiv.org/pdf/2004.03329.pdf. U okviru naučnog rada naveden je dataset MedDialog za kog tvorci tvrde da je trenutno najveći korpus dijaloga između pacijenata i lekara. Dataset sadrži 1,1 milion dijaloga na kineskom kao i 0.26 miliona dijaloga na engleskom jeziku. Pošto ćemo se u našem radu fokusirati samo na engleski jezik, iskoristićemo 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 se objašnjava zdravstveno stanje pacijenta i dela razgovara pacijenata i lekara. Podaci su preuzeti sa dva sajta - *icliniq.com* i *healthcaremagic.com*, koji predstavljaju online platforme za davanje stručnih medicinskih saveta. Podaci su dati u obliku tekstualnih datoteka. Do podataka je moguće doći preko https://github.com/UCSD-AI4H/Medical-Dialogue-System a omogućen je i direktan pristup tekstualnim datotekama preko https://drive.google.com/drive/folders/1g29ssimdZ6JzTST6Y8g6h-ogUNReBtJD?usp=sharing.

Dataset pronađen preko *Kaggle* sajta nazvan *Diagnose me* predstavlja json format prethodno navedenih podataka. Zbog veće mogućnosti rukovanja podacima koji se nalaze u json format, odlučili smo da iskoristimo ovaj dataset. Do podataka je moguće doći preko https://www.kaggle.com/datasets/dsxavier/diagnoise-me.

Drugi izvor podataka, odnosno skup dijaloga, pronađen je pretraživanjem *Github.com* sajta. Ovaj dataset je formiran na osnovu podataka sa sledećih veb sajtova - *eHealth Forum, iCliniq, Question Doctors* i *WebMD*. Navedeni podaci nalaze se u json formatu. Link do podataka je https://github.com/LasseRegin/medical-question-answer-data.

5. Predložena metodologija

Najpre će se vršiti pretpocesiranje podataka, a zatim transformacija podataka primenom NLP tehnika. Koristićemo Recurrent Neural Network (RNN), Sequence To Sequence model. Sequence To Sequence model se sastoji iz enkodera i dekodera koji će biti implementirani korišćenjem LSTM (Long short-term memory).

6. Metod evaluacije

Nakon čitanja istraživačkog rada na temu evaluacije chatbota (https://www.jmir.org/2020/6/e18301/PDF), došli smo do zaključka da je način evaluacije jedna od najvećih prepreka u napretku chatbot tehnologije. Zbog same prirode konverzacijskih chatbota, jedini načini evaluacije su manelno testiranje i procena rada chatbota na osnovu utisaka korisnika. Upravo zbog takvog načina evaluacije, uticaj subjektivnosti može poremetiti realne rezultate koje postiže chatbot. Iz tih razloga, ali i zbog primenjenih načina evaluacija u naučnim radovima navedenih u literaturi, opredelili smo se za sledeće načine evaluacije. Ti načini su:

- Korištenje CUQ (Chatbot Usability Questionary) alata koji je sličan SUS (System usabailty scale) alatu ali je specifično namenjen za evaluaciju chatbot-a. Detalji oko samog postupka izvršavanja ove vrste evaluacije nalaze se na sledećem linku https://www.ulster.ac.uk/__data/assets/pdf_file/0009/478809/Chatbot-Usability-Questionnaire.pdf.
- Korištenje BLEU (Bilingual Evaluation Understudy) rezultata, koji je u skorije vreme postao tipična metrika za evaluaciju chatbot usluga. Zbog korištenja ove metrike potrebno je da ulazne podatke podelimo na train i test skup.

7. Plan

- Prikupljanje podataka
- Pretprocesiranje i obrada podataka (primena NLP tehnika)
- Kreiranje modela
- Evaluacija modela i prilagođavanje parametara
- Analiza dobijenih rezultata

8. Tim

Teodora Maruna E2 63/2022, Vladimir Jovanović E2 60/2022