Зборник радова Факултета техничких наука, Нови Сад

UDK: 4.41
DOI: https://doi.org/10.24867/30BE03Blagojevic

ВИРТУЕЛНИ АСИСТЕНТ БАЗИРАН НА РАДУ ТРАНСФОРМЕР МОДЕЛА УЗ КОРИШЋЕЊЕ ВЕКТОРСКЕ БАЗЕ ПОДАТАКА

A VIRTUAL ASSISTANT BASED ON TRANSFORMER MODEL USING A VECTOR DATABASE

Његош Благојевић, *Факултет техничких наука, Нови Сад*

Област – УПРАВЉАЊЕ ДИГИТАЛНИМ ДОКУМЕНТИМА

Кратак садржај – *У овом раду биће приказана имплементација виртуелног асистента базираном на раду трансформер модела уз коришћење векторске базе података и напредних техника вештачке интелигенције. Систем ће бити специјализован на одређеном скупу података везаних за Факултет техничких наука у Новом Саду. Циљ система је олакшавање долажења до информација о факултету студентима који желе да упишу техничке науке.*

Кључне речи: *Семантичка претрага докумената, Трансформер модели, Виртуелни асистенти, Обрада природног језика.*

Abstract – *This paper presents the implementation of a virtual assistant based on transformer model using a vector database and advanced artificial intelligence techniques. The system will be specialized in a specific dataset related to the Faculty of Technical Sciences in Novi Sad. The goal of the system is to make it easier for students who want to enroll in technical sciences to get information about the faculty.*

Keywords: *Semantic search, transformer models, virtual assistants, natural language processing*

1. УВОД

Брз развој вештачке интелигенције и алгоритама машинског учења довео је до тога да технологија тренутно може, у мањој или већој мери, да замени човека у одређеним областима. Све више се развијају система који могу обрадити велике количине података и пружити корисне информације у реалном времену чиме се значајно повећава продуктивност и ефикасност.

У овом раду биће представљено решење које се базира на семантичкој претрази докумената које има за циљ добављање релевантних докумената за дати упит у реалном времену. Основна идеја је да се коришћењем модерних техника вештачке интелиген-ције, као што су трансформер модели, омогући пре-цизно и брзо проналажење информација у великом корпусу текстуалних података.

Конкретно, биће приказано решење које представља виртуелног асистента специјализованог на подацима везаним за Факултет техничких наука у Новом Саду. Овај алат ради на принципу семантичке претраге докумената где се документи чувају у виду вектора у векторској бази података. Упити корисника се такође преводе у векторе и на основу сличности се добављају релевантни подаци.

У овом раду најпре ће бити извршен преглед стања у области уз осврт на актуелна најсавременија решења. Потом ће бити описана архитектура и спецификација система након којих ће бити представљени најзначајнији имплементациони детаљи. Потом ће бити представљени и дискутовани резултати мерења перформанси алата у односу на друге, традиционалне начине претраге докумената. На крају, биће представљени закључци рада уз дискусију о могућим правцима даљег развоја.

2. ПРЕГЛЕД СТАЊА У ОБЛАСТИ

Обрада природног језика (*енг natural language processing*) је област вештачке интелигенције и лингвистике која се бави проучавањем проблема аутоматског производњења и разумевања природних људских језика [1].

Значајну прекретницу у развоју области обраде природног језика представља и настанак модела машинског учења од којих су најбитнији трансформер модели [2]. Примери ових модела јесу ”*BERT*”, “*GPT-3*” и разне варијанте настале од ових модела. Тренсформери су нашли примену у разним областима, од семантичке претраге, аутоматског превођења али и виртуелних асистената.

2.1. Виртуелни асистенти

Виртуелни асистенти представљају посебне алате и софтвере који могу да обављају низ задатака или услуга за корисника на основу корисничког уноса кроз текстуалне, али и вербалне команде.

Да развој ових софтвера представља примамљиво поље истраживања говори и чињеница да готово већина великих компанија из сфере софтверских информација улажу огроман новац у развој сопствених решења, тако да су неки од најпопуларнијих примера оваквих агената “*SIRI*” компаније “*Apple*”, амазонова “*Alexa*”, “*Google Assistant*” и “*Bixby*” компаније *Samsung*.

НАПОМЕНА:
Овај рад је проистекао из мастер рада чији ментор је био др Драган Ивановић, ред. проф.

2.2. *ChatGPT*

ChatGPT је модел заснован на машинском учењу који користи дубоке невронске мреже за генерисање текста [3]. Овај модел је трениран на великом обиму текстуалних података из различитих извора, укључујући интернет, књиге и новинске чланке. Његова основна функција је да одговара на корисничке упите и успостави дијалог са њим.

Шта чини *ChatGPT* тако напредним је његова способност да производи природан и разнообразан текст који одговара на различите захтеве корисника. Напредни алгоритми обраде природног језика омогућавају разумевање контекста упита и генерисање релевантних одговора, често са изненађујућом тачношћу.

3. КОРИШЋЕНЕ ТЕХНИКЕ И ТЕХНОЛОГИЈЕ

Серверска страна апликације развијана је уз коришћење *Flask* радног окружења у *Python* програмском језику.

Клијентска страна система је развијена уз помоћ *React* библиотеке у *JavaScript* програмском језику.

Као што је већ напоменуто, за складиштење података је коришћена векторска база података. Векторске базе података су тип базе података која складишти податке у векторском формату, што омогућава ефикасно извршавање упита и манипулацију огромним скуповима података [4]. За овај систем коришћена је *Pinecone* база података.

Обзиром да систем ради са великим бројем захтева који представљају питања корисника која се често понављају, наопходно је оптимизовати систем коришћењем технике кеширања. *Redis* представља брзу и флексибилну *in-memory* базу података намењену најчешће за кеширање података. Карактерише је брз приступ подацима и решавање проблема складишта малих количина података.

3.1. Трансформер модели

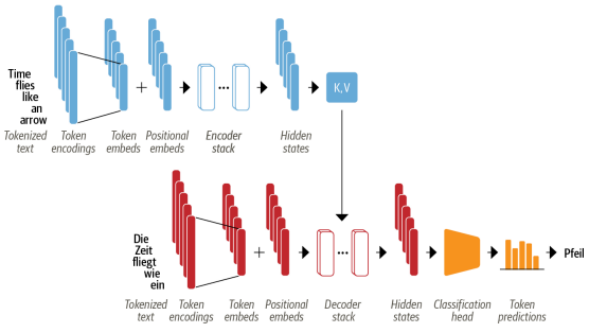
Трансформер модели представљају моделе дубоког учења развијене од стране *Google-a* 2017. године. Тада су представљени као замена за тадашње доминантне моделе трансформисања који су се ослањали на комплексне рекурентне или конволуционе неуронске мреже које укључују енкодер и декодер архитектуру кроз механизам пажње [5]. Трансформери су предложени као једноставније решење које ће се ослањати само на механизам пажње.

За разлику од традиционалних енкодер-декодер модела који су посматрали улазну секвенцу униформно (обрађиван је елемент по елемент), механизам пажње омогућава селективни приступ различитим деловина улазних података додељујући различите степене битности [6].

Механизам пажње је првобитно имплементиран у оквиру машинског превођења кроз модел назван *RNNsearch*. Овај модел користи двосмерну рекурентну неуронску мрежу као енкодер и декодер који опонаша претраживање кроз изворну реченицу приликом генерисања превода [6].

Енкодер израчунава скривена стања (анотације) за сваку реч у улазној секвенци, док декодер користи блок пажње за динамичко одређивање који делови улазне секвенце су најважнији за генерисање сваког следећег симбола у излазу [7].

На следећој слици је представљена архитектура енкодер-декодер трансформера узета из овог рада [7].



Слика 1. *Архитектура*

Оваква трансформер архитектура је оригинално била дизајнирана за *sequence to sequence* задатке попут машинског превођења. Међутим, временом су и енкодер и декодер постали посебне компоненте тако да сада постоји више трансформер модела базирано на томе које компоненте садржи.

3.1. Модели за генерисање одговора

Модели за одговор на питања (*QA* модели) су једна од најпопуларнијих области у обради природних језика. Ови модели имају способност да разумеју и генеришу одговоре на питања постављена на природном језику, што их чини изузетно корисним у бројним апликацијама, укључујући виртуелне асистенте, претраживаче и системе за подршку корисницима.

QA модели функционишу тако што обрађују улазни текст, који се састоји од пасуса (контекста) и питања, и затим генеришу одговор који је обично фраза или реченица из контекста. Уопштено, ови модели могу бити класификовани у две категорије:

- Екстрактивни модели
- Генеративни модели

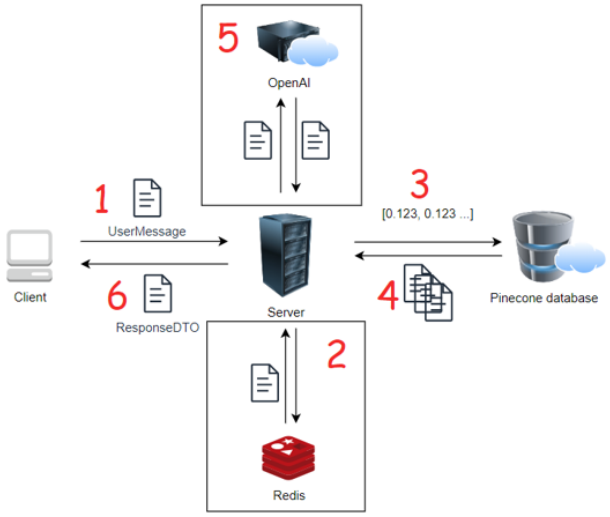
У оквиру овог рада нису имплементирани ови модели него се рад система ослања на коришћење *OpenAI* интерфејса и њиховог модела за генерисање одговора.

4. СПЕЦИФИКАЦИЈА

Систем се састоји од два главна дела, клијентске апликације са корисничким интерфејсом и серверске апликације. Серверска апликација комуницира са базом података преко *Pinecone* интерфејса али и користи *OpenAI* интерфејс за генерисање одговора на основу докумената добијених из базе података.

На следећој слици је приказана архитектура система са свим његовим модулима и приказаним корацима у

извршавању главног корисничког случаја, комуникације са асистентом.



Слика 2. *Архитектура и кораци извршавања*

У оквиру корака један, корисник уноси питање и шаље се захтев серверу. Питање је произвољног облика, може бити или реч или реченица. Пошто је могуће да се питања понављају проверава се у оквиру *Redis-a* да ли је дато питање већ било и враћа се кеширан одговор. Тиме се додатно оптимизује извршавање. Уколико то није сличај, питање се векторизује коришћењем истог модела који је коришћен за креирање базе података. Тај вектор се шаље бази за претрагу. База враћа листу од три најрелевантнија документа. Контексти се извлаче из резултата и шаље се упит преко *OpenAI* интерфејса за генерисање одговора на основу контекста. Одговор се шаље клијентској страни заједно за контекстима од кога је настао одговор.

5. ИМПЛЕМЕНТАЦИЈА

5.1. Скуп података

За задовољавајући рад трансформер модела неопходно је постојање добро припремљеног скупа података на коме би се модел могао тренирати. Пронађен је скуп података који је описан у раду [8] и који је синтетички направљен превођењем *SQUAD* скупа података са енглеског језика.

SQUAD скуп садржи око 87 хиљада елемената скупљених са разних извора, најчешће артикли са *Wikipedia*-е заједно са питањима везаним за тај артикал који су ручно формирани од стране великог броја људи.

Коришћени скуп података је дакле преведен скуп података на српски језик коришћењем *Translate-Align-Retrieve* методе описане у овом раду [21]. Начин на који је искоришћен скуп података полази од претпоставке да ће систем радити са питањима о факултету и контекстима који су издвојени део текста са сајта факултета или било ког другог извора, тако да се може искористити контекст и питање из *SQUAD* скупа података. Да би се трансформер модел могао тренирати потребно је да постоје нумеричке лабеле које одређују сличност између контекста и питања. Пошто у овом скупу лабеле не постоје, синтетички су креиране коришћењем више трансформер модела над скупом података на енглеском језику. Тиме се добијају релевантне лабеле које су касније коришћене у комбинацији са скупом података на српском језику.

5.2. Модел

Модел који је коришћен као основа за *finetuning* је *sentence-transformers_paraphrase-multilingual-MiniLM-L12-v2* . Овај модел мапира реченице и параграфе на векторе са 384 елемента. Модел се састоји од 118 милиона параметара и трениран је на 50 различитих језика, укључујући и српски.

Разлог зашто је изабран овај модел је управо тај што је у основи трениран и за српски језик. Тиме се у његовом речнику налазе и српске речи и способан је да ради и са текстом на српском језику. То је било неопходно да би се модел могао накнадно претренирати над скупом података на српском језику.

Скуп података је подељен на тренинг и валидациони део у односу 80-20. Функција губитка која је коришћена је *cosine similarity loss* а евалуатор је *embedding similarity evaluator*.

Batch size је подешен на 16 а број епоха је износио 7.

5.3. База података

Да би систем могао да функционише неопходно је имати базу података са релевантним информацијама о факултету. Најбољи извор информација је био званични сајт факултета тако да је већина информација прикупљена са сајта методом *web scraping*-а као текст директно из *HTML*-а. Такође, ручно су додате и информације са осталих извора као што су групе за упис на факултет које садрже одговоре на разна питања о факултету па су послужили као идеалан скуп података за овај систем. Једна од таквих група је и *Facebook* група „Бићу студент ФТН – упис 2024“ са најрелевантнијим подацима везаним за факултет.

6. РЕЗУЛТАТИ

Када се говори о резултатима модела трансформера, јако је тешко нумерички одредити прецизност алата управо због, како је већ наведено, синтетичког генерисања скупа података над којим је модел трениран. Чак и да то није случај, додатну проблематику представља чињеница да сличност два текста зависи и од самих субјеката који раде лабелирање, а потпуна униформност приликом лабелирања је готово немогућа.

С тим у вези, иако ће се приказати евалуациони резултати добијени приликом тренирања, ово поглавље ће бити више посвећено упоређивању овог приступа са класичним претрагама докумената коришћењем *elastic search-a*. Што се тиче евалуационих резултата модел евалуиран коришћењем различитих метричких приступа. Након тренирања модел је показивао тачност од 0.843 у

косинусној сличности, 0.774 у еуклидској и 0.763 у менхетенској. Међутим, најбоља евалуација оваквих модела је тестирање и упоређивање са системима који користе другачији приступ за претрагу докумената.

Систем је упоређиван са класичним системом за претрагу докумената користећи упите различитих дужина. Истраживање је показало да се класични системи показују боље само када су упити доста кратки и када немамо довољно контекста за семантичку претрагу. Тако да ситуације у којима ће класични системи боље радити јесу претраге по кључним речима или по одређеним фразама. Повећавањем упита и проширивањем контекста систем базиран на трансформер моделу показује боље резултате док највећу предност показује у ситуацијама када се упит састоји од речи које не постоје у документу већ су само семантички сличне. У таквим ситуацијама класични системи постају готово неупотребљиви.

7. ЗАКЉУЧАК

У овом раду је представљен један начин имплементације виртуелног асистента специфицираног на одређен домен знања (у овом случају информације се примарно добављају са сајта Факултета техничких наука у Новом Саду). Циљ оваквог система је приближавање информација факултета техничких наука корисницима а све у циљу повећања доступности информација о самом факултету. Имплементација решења је обухватала различите кораке и делове система који се могу посматрати независно.

Први корак у имплементацији је прикупљање базе знања. За овај конкретан случај, документи за базу су прикупљени са сајта методом *web scraping*-а.

Други део система је систем за претрагу докумената базиран на трасформер моделима. Конкретно за овај систем је коришћен претраниран модел *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* који је дотрениран подацима на српском језику.

Трећи део јесте генерисање одговора на основу упита и резултата. Због недостатка података за тренирање генеративних модела за одговарање, коришћен је *OpenAI* интерфејс којем су прослеђивани резултати добављени у претходној фази.

Што се тиче корисничког дела, апликација се може користити као виртуелни асистент у оквиру сајта факултета, као посебна апликација за дописивање и постављање питања али и кроз административни режим за лаку манипулацију подацима из базе података.

Када говоримо о начинима надоградње система, могућности су разне. Полазећи од добављања података за базу који се могу аутоматизовати и чији извор може бити, како сајт, тако и приватан архив факултета. Са повећањем докумената у бази повећава се и спектар могућих питања и одговора са којима систем може да ради.

Трансформер модели се развијају свакодневно тако да ће вероватно у скорије време бити унапређени и вишејежични модели који обухватају и српски језик пошто у овом тренутку само мало број модела ради са српским језиком а и скупови података за тренирање готово да ни не постоје на српском језику. Унапређењем овог дела би се унапредила и претрага докумената тако да би увек имали најрелевантније документе припремљене за наредну фазу.

Последња фаза уједно оставља и највише простора за напредовањем где би циљ био креирање сопственог модела за одговарање на питање у слободној форми. Такође, модел би морао да буде способан за рад са документима на српском језику.

8. LITERATURA

[1] Chowdhary, K.R. (2020). Natural Language Processing. In: Fundamentals of Artificial Intelligence. Springer, New Delhi. https://doi.org/10.1007/978-81 322-3972-7_19

[2] A. Gillioz, J. Casas, E. Mugellini and O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," 2020 15th Conference on Computer Science and Information Systems (FedCSIS), Sofia, Bulgaria, 2020, pp. 179-183, doi: 10.15439/2020F20.

[3] An, J., W. Ding, and C. Lin. "ChatGPT." tackle the growing carbon footprint of generative AI 615 (2023): 586.

[4] Han, Yikun, Chunjiang Liu, and Pengfei Wang. "A comprehensive survey on vector database: Storage and retrieval technique, challenge." arXiv preprint arXiv:2310.11703 (2023).

[5] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

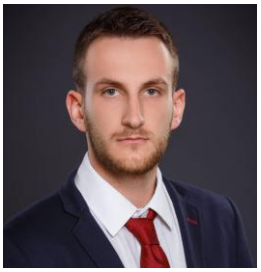
[6] Niu, Zhaoyang, Guoqiang Zhong, and Hui Yu. "A review on the attention mechanism of deep learning." Neurocomputing 452 (2021): 48-62.

[7] Tunstall, Lewis, Leandro Von Werra, and Thomas Wolf. Natural language processing with transformers. " O'Reilly Media, Inc.", 2022.

[8] Cvetanović, Aleksa, and Predrag Tadić. "Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian." 2023 31st Telecommunications Forum (TELFOR). IEEE, 2023.

[9] Carrino, Casimiro Pio, Marta R. Costa-Jussà, and José AR Fonollosa. "Automatic spanish translation of the squad dataset for multilingual question answering." *arXiv preprint arXiv:1912.05200* (2019).

Кратка биографија:



Његош Благојевић рођен је 29. 04. 1999. у Зворнику. Смер софтверско инжењерство и информационе технологиј на Факултету техничких наука у Новом Саду уписао је 2018. године. Основне студије је завршио у септембру 2022.године и од октобра 2022. године студира на мастер студијама Факултета техничких наука.