Зборник радова Факултета техничких наука, Нови Сад

UDK: 4.8
DOI: https://doi.org/10.24867/32BE19Dragicevic

ДИГИТАЛНИ АСИСТЕНТ ЗА ОДГОВАРАЊЕ НА ПИТАЊА У ДОМЕНУ ОСИГУРАЊА

INSURANCE DOMAIN QUESTION-ANSWERING DIGITAL ASSISTANT

Стефан Драгичевић, Јелена Сливка, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИЧКО И РАЧУНАРСКО ИНЖЕЊЕРСТВО

Кратак садржај – *У овом раду представљен је развој напредног четбот система намењеног за аутоматизовано одговарање на питања из области осигурања. Систем је имплементиран као веб апликација која омогућава корисницима брз и јасан одговор на постављено питање. Систем се састоји од неколико кључних модула: модула за претпроцесирање упита, модела за проналажење релевантних питања (модел високог одзива), модела за филтрирање и одабир најрелевантнијег одговора (модел високе прецизности), модула за класификацију сентимента одговора и модула за издвајање кључних речи.*

Кључне речи: *Процесирање природног језика, дигитални асистент за конверзацију, класификација сентимента*

Abstract – *This paper presents the development of an advanced chatbot system designed to answer insurance-related questions automatically. The system is implemented as a web application that provides users with quick and clear answers to their questions. The system consists of several key modules: a query preprocessing module, a model for finding relevant questions (high recall), a model for filtering and selecting the most relevant answer (high precision), a sentiment classification module, and a keyword extraction module.*

Keywords: *NLP, chatbot, LSTM, sentiment classification*

1. УВОД

У последњих неколико година, дигитални асистенти за конверзацију, односно, четботови (*chatbots*), су постали кључни део дигиталне трансформације и интеракције са корисницима [1]. Са растућим утицајем вештачке интелигенције (ВИ) и аутоматизације процеса, четботови су постали снажан алат за компаније и организације широм света. Ови дигитални асистенти омогућавају интеракцију са корисницима кроз текстуалне или гласовне команде, нудећи одговоре на упите, решавајући проблеме и пружајући корисничку подршку у реалном времену.

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је била др Јелена Сливка, ванр. проф.

Једна од кључних предности четботова лежи у њиховој сталној доступности. Независно од временске зоне или локације, корисници могу добити тренутне одговоре на своја питања или решити недоумице у било ком тренутку. Овај рад се фокусира на развој четбота који је намењен одговарању на питања из области осигурања који, поред тога, има и способност препознавања кључних речи одговора, као и могућност анализирања сентимента одговора. Овакав четбот може бити имплементиран на разним веб сајтовима како би побољшао корисничко искуство и смањио оптерећење корисничког центра.

Четбот се састоји од више независних модула:

1. Модул за претпроцесирање упита

2. Модел који враћа сва питања која би могла бити релевантна за дати упит. Битна особина овог модела је да има велики одзив (*high recall*).

3. Модел који филтрира одговоре које је вратио *high recall* модел. Битна особина овог модела је да има високу прецизност (*high precision*) како би од свих релевантних одговора вратио одговор који најбоље одговара задатом упиту.

4. Модел за класификацију сентимента пронађеног одговора.

5. Модел за проналажење кључних речи у пронађеном одговору.

Сваки модел је обучаван независно, на различитим скуповима података, и евалуиран метрикама које одговарају датом задатку. Сви модели дају прилично добре резултате, што омогућава високе перформансе целокупне апликације. Поређења ради, најбољи модел за проналажење најсличнијег питања на *benchmark* тестовима [7] има тачност од 92,3% док *LSTM (Long short-term memory)* модел који је коришћен у овом раду има тачност од 89% на скупу података за тестирање. Директно поређење резултата које дају модели у овом раду са онима из литературе није могуће јер овај рад користи тест скупове података који се разликују од оних из литературе.

2. ПРЕГЛЕД СТАЊА У ОБЛАСТИ

Овај рад се бави решавањем више мањих проблема повезаних са сличношћу реченица, класификацијом сентимента и издвајањем кључних речи. Због тога ће ово поглавље бити организовано у потпоглавља која обрађују сваки од ових проблема.

2.1. Сличност речи и реченица

У раду [2] предложена је употреба *Word2Vec* методе за одређивање сличности између речи. Свака реч се трансформише у вектор унапред дефинисане дужине. Током тренинга *Word2Vec* анализира велике корпусе текста и учи да предвиђа речи које се појављују у околини циљане речи (*skip-gram*) или обрнуто, да предвиђа циљану реч на основу околних речи (*continuous bag of words*). Овакав приступ је ефикасан јер семантички сродне речи заузимају блиске позиције у векторском простору. Модел је обучаван на википедији, при чему је за поређење речи коришћено косинусно растојање. За евалуацију се користио Пирсонов коефицијент корелације, чије вредности варирају између -1 и 1, где вредност 1 указује на потпуну корелацију.

Рад [3] предлаже да се за одређивање сличности између реченица користи сијамски *LSTM* модел. За обучавање тог модела је коришћен SICK скуп података који се састоји од 10 000 парова реченица на енглеском језику. Сваки пар реченица има оцену сличности од 1 до 5 које је добијена усредњавањем процене 10 различитих особа. За евалуацију модела је коришћена Пирсонова корелација. Будући да је модел из студије [3] евалуиран на другачијем скупу података, директна компарација резултата тог истраживања са налазима овог рада није могућа. Аутори студије [3] наводе да њихов модел постиже средњу квадратну грешку од 0.5093, што сматрају задовољавајућим резултатом.

2.2. Класификација сентимента

Сентимент реченице може се утврдити применом различитих приступа. Они укључују лексичке методе, технике базиране на дефинисању правила и методологије машинског учења. У раду [4] су описане предности и мане сваке од ових метода. Рад [4] не наводи скуп података који је коришћен за обучавање и евалуацију модела, али је закључено да методе засноване на машинском учењу дају боље резултате у односу на остале методе које су испробане.

Рад [5] пореди најпопуларније библиотеке за класификацију сентимента које користе лексички приступ. Међу анализираним библиотекама налазе се NLTK, *Vader* и *Textblob*. За евалуацију је употребљен скуп који обухвата 11 861 реченицу на енглеском језику, класификованих као позитивне или негативне. Најбоље перформансе остварила је *Vader* библиотека са прецизношћу од 77%, док је *Textblob* показао нешто слабије резултате са тачношћу од 74%. Одзив и Ф-мера ових приступа показују сличне вредности, око 80%.

2.3. Издвајање кључних речи

Количина података доступних на интернету експоненцијално расте. За ефикасну организацију и препоручивање овог садржаја, неопходна је идентификација кључних термина. Рад [6] предлаже да се за решавање проблема одређивања кључних речи користи *TF-IDF*. У раду [6] се користи ненадгледано учење јер је анотирање кључних речи напорно и временски захтевно, док резултати могу бити субјективни. За овај експеримент су у раду [6] коришћена два скупа података, *News* и *Computer*. За евалуацију резултата су аутори користили прецизност, одзив и Ф-меру. Анализа је показала значајну зависност резултата од параметра који дефинише број селектованих кључних речи, са варијацијама које достижу и 10%. Упркос томе, истраживање демонстрира да предложена методологија постиже задовољавајуће резултате.

3. МЕТОД

У овом поглављу је представљена имплементација четбота који одговара на питања из осигурања. Улаз у овај систем представља питање које је корисник поставио, а излаз из система је одговор на најсличније питање из базе, његов сентимент и наглашене кључне речи. Имплементација се састоји из дела за претпроцесирање (поглавље 3.1), *high recall* дела (поглавље 3.2), *high precision* дела (поглавље 3.3), као и модула за класификацију сентимента (поглавље 3.4) и модула за издвајање кључних речи (поглавље 3.5).

3.1. Претпроцесирање

Претпроцесирање текста представља есенцијалну фазу у припреми података за моделе машинског учења и обраде природног језика. Процес трансформише корисничко питање у векторску репрезентацију погодну за даљу анализу. Кључне фазе обухватају:

1. Елиминацију стоп речи и специјалних карактера који не доприносе значењу текста.

2. Нормализацију текста конверзијом у мала слова.

3. Лематизацију, која своди речи на њихов речнички облик.

4. *Stemming* технику, која је тестирана али није имплементирана због инфериорних резултата.

5. Векторизацију помоћу *word2vec* модела.

6. Генерисање вектора реченице сумирањем појединачних вектора речи, што је метода која показује добру ефикасност у пракси [16].

Идентичан процес се примењује на сва питања у бази података, обезбеђујући конзистентност у обради и прецизно поређење упита.

3.2. *High recall* део: враћање свих могућих одговора

Ова фаза је кључна у системима за одговарање на питања, јер омогућава проналажење релевантних информација из постојеће базе знања. Улаз у овај део је векторизовано питање, које представља резултат претходног корака претпроцесирања. Излаз је скуп најсличнијих питања из базе, који ће се даље користити за генерисање одговора. Процес проналажења најсличнијих питања укључује следеће кораке:

1. Поређење векторизованог питања са свим питањима из базе. Поређење се врши у векторском простору, где свако питање представља тачку у том простору.

2. Примена *KNN* (*K-Nearest Neighbors*) модела за проналажење најсличнијих питања

3. Одабир метрике сличности. У коначној верзији система, *KNN* модел користи Менхетн метрику за мерење удаљености између векторизованих питања јер су експерименти показали да ова метрика даје најбоље резултате у овом конкретном случају.

4. *KNN* модел проналази 150 питања из базе која су најсличнија постављеном питању према Менхетн метрици.

Овај приступ омогућава ефикасно претраживање велике базе питања и проналажење оних која су најсличнија корисничком упиту.

3.3. *High precision* део: филтрирање добијених одговора

Циљ овог дела је да пронађе питање из претходно издвојеног скупа које је најсличније корисничком упиту. Улаз у овај део чине 150 издвојених питања из претходног корака, као и оригинално питање које је поставио корисник. Излаз је одговор на најсличније питање пронађено у бази.

За поређење сличности питања и одговора је коришћена сијамска *LSTM* неуронска мрежа дизајнирана за мерење сличности између парова секвенци. Сијамска мрежа се састоји од две идентичне подмреже које деле исте параметре. Свака подмрежа обрађује по једно питање, а затим се њихови излази комбинују како би се израчунала мера сличности.

Сијамска *LSTM* мрежа је претходно тренирана на великом скупу парова питања, где је учила да разликује слична од различитих питања. Ово тренирање омогућава мрежи да ефикасно мери семантичку сличност између питања, чак и када она нису формулисана на идентичан начин.

За свако од 150 питања из базе, сијамска *LSTM* мрежа генерише скор који представља степен сличности датог питања са корисничким питањем. Овај скор је нумеричка вредност која квантификује колико су два питања семантички блиска. Сва питања из базе се рангирају на основу добијених скорова сличности, а питање са највишим скором се сматра најсличнијим корисничком упиту. Одговор који је повезан са изабраним најсличнијим питањем у бази се узима као коначан одговор на корисничко питање.

3.4. Модул за класификацију сентимента

Након проналажења одговора на корисничко питање, тај одговор се прослеђује ансамблу модела који класификују сентимент, а који се састоји од *XGBoost, SVM* (*Support Vector Machines*) модела и лексичког приступа. Излаз из сваког модела се множи одговарајућим тежинама. Тежине су: 0.4 за *XGBoost* модел, 0.4 за *SVM* модел и 0.2 за лексички модел. Ове тежине су одабране јер *XGBoost* и *SVM* модели показују знатно боље перформансе у поређењу са лексичким моделом, те, у случају супротстављених предикција ових модела, коначну одлуку доноси лексички модел. Излаз из овог модула је број у опсегу од 0 до 1, при чему се свака вредност већа од 0.5 сматра као позитиван сентимент.

3.5. Модул за издвајање кључних речи

Улаз у овај модул је одговор на питање из кога се извлаче кључне речи. Процес започиње анализом текста коришћењем техника обраде природног језика. Означавају се делови говора (*POS tagging*), где се свака реч у тексту класификује према својој граматичкој улози, као што су именице, глаголи и придеви.

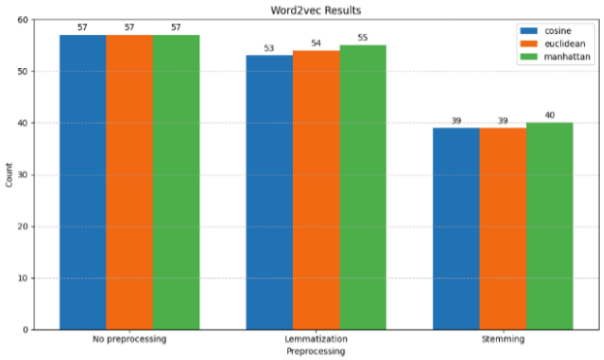
У овом случају, фокус је на именицама једнине и множине, као и на властитим именицама. Именице се детектују помоћу NLTK библиотеке. Након идентификације именица, оне се рангирају помоћу *TF-IDF* методе. Из рангиране листе именица се извлачи пет најзначајнијих, које се сматрају кључним речима текста. Ове речи представљају суштинске концепте и теме обрађене у анализираном одговору. Одабран је приказ пет кључних речи омогућава јер се тако омогућава брз увид у главне теме текста, без преоптерећивања корисника превеликим бројем информација.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

За потребе овог истраживања коришћена су три главна скупа података. Скуп података који је коришћен за тренирање четбота да препозна најсличније питање је *Quora dataset* [8]. Овај скуп података служи за тренирање *LSTM*-а како би могао да пронађе најсличније питање из базе. Други скуп података је IMDB скуп података [9], који се користи за класификацију сентимента. Трећи скуп података је *insurenceQA* [10] који се састоји од парова питања и одговора из области осигурања. Он је коришћен као база података за читав пројекат, односно из њега се тражи најсличније питање са питањем које је поставио корисник и даје се одговор на то питање.

Циљ првог експеримента је био да се одреди оптимална конфигурација *KNN* модела за проналажење најсличнијих питања. Тестирани су различите величине вектора речи, различите метрике удаљености: Еуклидска, косинусна и Менхетн метрика као и различите методе за векторизацију речи *TF-IDF* и *word2vec*.

На слици 1 су приказани резултати који се добијају када се користи *word2vec* метода за векторизацију речи. Најбољи резултати се добијају када се не врши препроцесирање текста, независно од коришћене метрике сличности. Овај приступ даје значајно боље резултате него *TF-IDF* метода, те је имплементиран у *high-recall* делу. На тест скупу, овај метод је тачно издвојио 54 од укупно 60 питања у скуп најсличнијих питања.



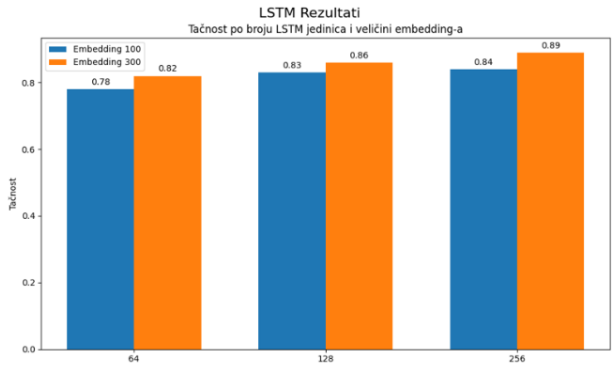
Слика 1. *Број питања која су се нашла у скупу најсличнијих питања у односу на број карактеристика, word2vec приступ*

Циљ другог експеримента је био да се оптимизује архитектура сијамске *LSTM* мреже за мерење сличности питања. Тестиране су различите вредности хипер-параметара за број *LSTM* јединица 64, 128, 256 и

величину *embedding* слоја 100, 300. Мрежа је тренирана на *Quora* скупу података који је подељен на тренинг, тест и валидациони у односу 70:15:15. Перформансе су мерене кроз тачност класификације сличних и различитих парова питања. Чува се модел који има најбољу тачност на валидационом скупу.

У експерименту је утврђено да повећање броја *LSTM* јединица доводи до побољшања перформанси (Слика 2). Највећи скок у перформансама се види при преласку са 64 на 128 јединица. Прелазак са 128 на 256 јединица доноси мање побољшање, што сугерише да се приближавамо тачки засићења.

Већи ембединг слој конзистентно доводи до бољих резултата, што се види на слици 2. Побољшање је приметно при повећању са 100 на 300. Ово сугерише да већи ембединг слој омогућава боље представљање семантичких карактеристика питања. Најбољи резултати су постигнути са 256 *LSTM* јединица и ембединг слојем величине 300.



Слика 2. *Тачност за различите вредности хипер-параметара LSTM модела*

У трећем експерименту упоређено је неколико модела машинског учења за класификацију сентимента користећи IMDB скуп података. Модели машинског учења су тренирани и евалуирани са различитим бројем обележја, односно, коришћен је *TF-IDF* вектор различите дужине: 100, 500, 1000 и 2000. Модели који су коришћени су: *SVM*, *KNN, XGBoost*, модел насумичне шуме *(Random Forest)* и неуронска мрежа. Са слике 3 се види да најбоље резултате даје *SVM* модел када је величина вектора речи 2000. Сличне резултате даје и *XGBoost* модел када је величина вектора 1000.

Слика 3. *Одређивање сентимента применом метода машинског учења*

У четвртом експерименту развијена је функција која комбинује *TF-IDF* методу са NLTK библиотеком за извлачење кључних речи. Процес укључује конверзију текста у мала слова, токенизацију, означавање врста речи (са фокусом на именице), и *TF-IDF* анализу која додељује веће тежине речима које су честе у датом тексту али ретке у општем корпусу. Ефикасност ове методе није могла бити процењена јер *insurenceQA* база података не садржи лабелиране кључне речи.

5. ЗАКЉУЧАК

Рад је представио дигиталног асистента за конверзацију који пружа одговоре на питања из области осигурања. Мотивација за његов развој лежи у растућој потреби за аутоматизованом корисничком подршком у индустрији осигурања, која може значајно побољшати корисничко искуство и смањити оптерећење корисничких центара. Систем користи комбинацију техника обраде природног језика и машинског учења. Система је евалуиран низом експеримената који су тестирали његове различите аспекте.

6. ЛИТЕРАТУРА

[1] Adamopoulou & Moussiades (2020). An Overview of Chatbot Technology. AIAI 2020, IFIP Advances, 584. Springer.

[2] Jatnikaa et al. (2019). Word2VecModel Analysis for Semantic Similarities. ICCSCI 2019.

[3] Lv et al. Siamese Multiplicative LSTM for Semantic Text Similarity.

[4] Devika et al. Sentiment Analysis: A Comparative Study On Different Approaches. ICRTCSE.

[5] Bonta et al. (2019). A Comprehensive Study on Lexicon-Based Approaches. Asian J Comp Sci Tech, 8(S2).

[6] Xu & Zhang (2020). Extracting Keywords from Texts. ICIIK.

[7] https://paperswithcode.com/sota/question-answering-on-quora-question-pairs

[8] https://huggingface.co/datasets/quora-competitions/quora

[9] https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[10] https://github.com/shuzi/insuranceQA

Кратка биографија:

Стефан Драгичевић је рођен 1998. године у Новом Саду, Србија. Основне академске студије завршио је 2022. године на Факултету техничких наука. Исте године уписује мастер за вештачку интелигенцију који завршава 2 године касније одбраном мастер рада из области обраде природног језика.
контакт: stefand998@gmail.com