Зборник радова Факултета техничких наука, Нови Сад

UDK: 004.42:004.738.12
DOI: https://doi.org/10.24867/33BE34Kanjuh

ОПТИМИЗАЦИЈА ЧЕТБОТА КОРИШЋЕЊЕМ ТРАНСФОРМЕР МОДЕЛА

OPTIMIZATION OF A CHATBOT USING TRANSFORMER MODELS

Нађа Кањух, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *Циљ рада је оптимизација конверзацијског четбота за домен осигурања коришћењем алгоритама природне обраде језика и трансформер модела, конкретно BERT модела, ради бољег разумевања језика и термина специфичних за осигурање. У истраживању је четбот обучен и тестиран у поређењу са LSTM моделом, при чему су експериментални резултати показали да BERT модел даје боље резултате због своје способности разумевања ширег контекста у упитима корисника. Ова способност омогућава четботу да прецизније интерпретира сложене и контекстуално богате упите, што је посебно важно за пружање тачних и поузданих информација у домену осигурања.*

Кључне речи: *четбот, трансформер модели, BERT, LSTM, природна обрада језика, велики језички модели*

Abstract – *The goal of this thesis is to optimize a conversational chatbot for the insurance domain using natural language processing algorithms and transformer model, specifically the BERT model, to enhance understanding of language and insurance-specific terminology. In the research, the chatbot was trained and tested in comparison with the LSTM model, with experimental results showing that the BERT model performs better due to its ability to understand broader context in user queries. This capability allows the chatbot to interpret complex and contextually rich queries more accurately, which is especially important for providing precise and reliable information in the insurance domain.*

Keywords: *chatbot, transformer models, BERT, LSTM, natural language processing – NLP, large language models*

1. УВОД

Имплементација четбота доноси низ предности. Првенствено, четботоби могу побољшати корисничко искуство пружањем брзих и тачних одговора на захтеве клијената, било да се ради о захтевима за информацијама, обради одређених или пружању подршке у реалном времену. Поред тога, они омогућавају уштеду ресурса и времена, аутоматизујући рутинске задатке и смањујући потребу за директним људским ангажовањем.

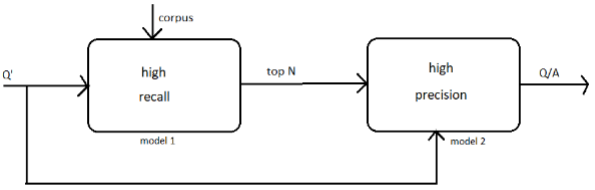
НАПОМЕНА:

Овај рад проистекао је из мастер рада чији је ментор био др Милан Сечујски, ред. професор.

С обзиром на то да *LSTM*, као врста рекурентних неуронских мрежа, често није у стању да ухвати довољно широк контекст, све је већи интерес за њихову замену ради унапређења перформанси конверзацијских система. У овом раду испитује се побољшање конверзацијског система применом трансформер модела, конкретно *BERT* модела. *BERT* модел пружа напредне функције кодирања са двосмерном пажњом, што омогућава боље разумевање значења речи у контексту целе реченице.

Циљ овог рада је да покаже како се систем може унапредити коришћењем савремених трансформер модела, као и да упореди резултате *LSTM* и *BERT* модела.

Иако савремени приступи попут *Retrieval-Augmented Generation – RAG* омогућавају четботовима да приступе обимним базама знања и генеришу ажурне одговоре, њихова примена може бити ограничена доступношћу квалитетно структуираних података и прихватљивим временом одзива. Због тога је у овом раду изабран приступ двостепеног одлучивања, који комбинује модел високог одзива (енг. *high recall model*) и модел високе прецизности (енг. *high precision model*). Модел високог одзива је дизајниран да идентификује што већи број релевантних кандидата за одговор на упит из корпуса. Углавном се за издвајање релевантних кандидата бирају статистички модели због своје брзине и ефикасности. Са друге стране, циљ модела високе прецизности јесте да идентификује најадекватније питање од свих издвојених најбољих кандидата и да се на тај начин кориснику пружи одговарајући одговор. У овом кораку често се користе архитектуре попут *LSTM* и *BERT* модела. На слици 1 се налази блок дијаграм двостепног одлучивања.



Слика 1. *Блок дијаграм двостепеног одлучивања*

2. ДЕФИНИЦИЈЕ

У овом сегменту је дато објашњење најбитнијих теоријских појмова и алгоритама који се користе за израду четбота.

2.1. Модел високог одзива

У наставку ће укратко бити описани основни кораци за имплементацију модела високог одзива, обрада текста и векторизација.

2.1.1. Стеминг

Стеминг (енг. *Stemming*) је процес обраде текста који подразумева свођење рели на основни или коренски облик. Циљ стеминга је поједностављивање речи уклањањем префикса или суфикса, али уз задржавање основног облика. Неке од предности стеминга су смањење речника, побољшање претраге и брже процесирање. Са друге стране неке од мана јесу потенцијално прекомерно редуковање, затим изазови стеминг методе за поједине језике, као свођење различитих речи на исти корен што често може довести до конфузије у задацима као што су генерисање текста.

2.1.2. Лематизација

Лематизација (енг. *Lemmatization*) је процес свођења речи на основни облик или облик из речника, познат као лема. За разлику од стеминга који једноставно уклања префиксе или суфиксе, лематизација узима у обзир морфолошку анализу речи и има за циљ одређивање њеног основног облика. Предности коришћења лематизације јесу боље задржавање семантичког значења, побољшање претраге и боље разумевање језика. Са друге стране, лематизација може бити поприлично сложена и захтева по питању ресурса, а такође може доћи и до губитка одређених морфолошких информација приликом свођења речи.

2.1.3. *N*-грами

N-грами представљају узастопне низове од *n* елемената одређеног скупа, при чему елемент може бити реч, карактер или чак фонема, у зависности од конкретног проблема. У контексту обраде природног језика, *n*-грами се најчешће односе на секвенце речи. *N*-грами бележе локални заједнички контекст речи и могу пружити корисне увиде у односе и обрасце унутар текста. Један од недостатака јесте тај да коришћење *n*-грама може довести до реткости података, а због фиксне величине прозора ограничени да ухвате шири контекст.

2.1.4. Учесталост термина

Учесталост термина (енг. *Term frequency – TF*), представља један од најједноставнијих начина векторизације текстуалних података. *TF* је основни концепт у обради природног језика који мери фреквенцију појављивања термина у документу. Када се примењује *TF* као метода векторизације, сваки документ се представља као вектор чија дужина одговара броју јединствених термина у целокупном корпусу. Сваки елемент вектора представља фреквенцију одговарајућег термина унутар датог документа. Овај начин векторизације је погодан због своје једноставности и флексибилности у примени, а такође је добар и за истицање значаја термина унутар документа. Нека ограничења ове методе јесу та што додељује веће тежине речима које се често појављују, а такође не узима у обзир шири контекст, па може доћи до губитка информација.

2.1.5. Учесталост термина – инверзна учесталост на нивоу документа

Учесталост термина - инверзна учесталост на нивоу документа (енг. *Term Frequency – Inverse Document Frequency, TF-IDF)* је метода векторизације текста која комбинује локалну фреквенцију термина (енг. *Term Frequency - TF*) са глобалном значајношћу термина у корпусу (енг. *Inverse Document Frequency – IDF*). *TF* компонента наглашава значајне термине унутар појединачног документа. Са друге стране, *IDF* компонента има улогу да смањи важност честих термина који се појављују у целом корпусу. Ова метода може бити погодна за идентификацију кључних речи у документу, а такође и смањује важност неинформативних термина који се често појављују у целом корпусу. Неки од недостатака ове методе јесте не узимање у обзир редоследа речи у документу, као и недостатак контекста и семантичког разумевања. Ова метода не узима у обзир сложеније језичке елементе попут синонима, вишезначности речи итд.

2.1.6. *Word2Vec*

Основни циљ *Word2Vec* методе јесте извлачење семантичких информација и других битних особина из речи. *Word2Vec* користи контекстне информације речи како би научио њихове векторске репрезентације. Идеја је да сличне речи буду представљене у векторском простору једна близу друге. *Word2Vec* ради по сличном принципу као аутокодери. Постоје два приступа рада са *Word2Vec* методом. Први јесте коришћење претренираног модела чиме се добија велики број речи, али недостатак је што често модел није специјализован за домен над којим се ради. Други приступ подразумева тренирање модела, чиме добијамо модел који је специјализован за одређени домен, али мана овог приступа јесте недовољно речи. Коришћење *Word2Vec* доводи до тога да вектори генерисани помоћу овог алгоритма имају својство да задрже семантичке односе између речи, такође омогућава ефикасно рачунање сличности речи помоћу метрика удаљености. Недостатак је тај што квалитет векторских репрезентација зависти од величине и квалитета корпуса, као и то што *Word2Vec* узима у обзир само локални контекст око речи.

2.2. Модел високе прецизности

У наставку ће бити описана архитектура трансформера, њихова предност у односу на рекурентне неуронске мреже, а такође ће бити и описан начин функционисања *BERT* модела.

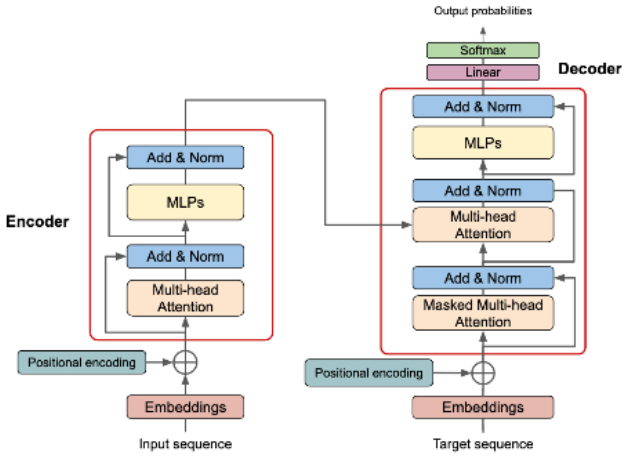
2.2.1. Предности трансформера у односу на рекурентне неуронске мреже

Трансформери (енг. *Transformers*) су један од најсавременијих и најнапреднијих модела вештачке интелигенције који омогућавају разумевање и генерисање текста на начин сличан човеку. Пре њих, рекурентне неуронске мреже су податке обрађивале

секвенцијално и то је довело до проблема споре обуке и губитка информација на дужим секвенцама. Трансформери решавају ове проблеме користећи механизам пажње, који омогућава моделу да истовремено обрађује све делове текста и фокусира се на кључне елементе без обзира на њихову удаљеност [1].

2.2.2. Архитектура трансформера

Када је у питању архитектура трансформера, они се састоје од два основна дела: кодера и декодера. Оба поменута дела се састоје од неколицине других слојева, а у наставку рада биће детаљно описан сваки део заједно са његовим слојевима. На слици 2 је приказана архитектура трансформер модела.



Слика 2. *Архитектура трансформера*

Један слој кодера (енг. *encoder*) се састоји од слоја вишеглавне пажње, два слоја нормализације, двослојне потпуно повезане мреже, а уз све то постоје и резидуалне конекције. Оваквих слојева има *N*. Улаз у кодер јесу векторисане речи са додатим позиционим ембединзима који су неопходни како би модел имао информацију о позицијама речи.

Механизам пажње (енг. *attention mechanism*) је главни елемент трансформер архитектуре. Он омогућава моделу да обради више пажње на део улазних података који садржи значајније информације и да посвети мање пажње остатку улаза [2]. У пракси, обично се не добијају добри резултати коришћењем једног слоја, па се из тог разлога обично израчунава више слојева пажње у паралели и резултати се конкатенирају. Вишеглавна пажња (енг. *Multi-head attention*) представља механизам који служи за поменуто паралелно израчунавање.

Слојеви нормализације помажу у одржавању стабилности током тренинга, смањују проблем нестајања или експлодирања градијента, помажу бољој генерализацији модела, а такође помажу моделу да се прилагоди различитим скалама улаза [2].

Резидуалне конекције имају кључну улогу у побољшању ефикасности тренирања и смањењу ризика од нестајања или експлодирања градијента у дубоким моделима. Оне омогућавају да све информације остану очуване током тренирања.

Потпуно повезана мрежа има за циљ да додатно обради податке и научи комплксне и нелинеарне везе између података. Она се углавном састоји од два скривена слоја и *ReLU* активационе фунцкије.

Излаз из слоја кодера је скуп вектора, при чему сваки представља улазни низ обогаћен контекстом. Овај излаз се користи као улаз у декодер трансформера.

Декодер (енг. *Decoder*) је практично исти кодер, осим што садржи додадни слој вишеглавне пажње која ради над излазом кодера. Циљ декодера јесте да комбинује излаз кодера са циљном секвенцоми прави предвиђања, односно да предвиди следећи токен. Број слојева декодера је углавном исти као и број слојева кодера.

2.2.3. Велики језички модели

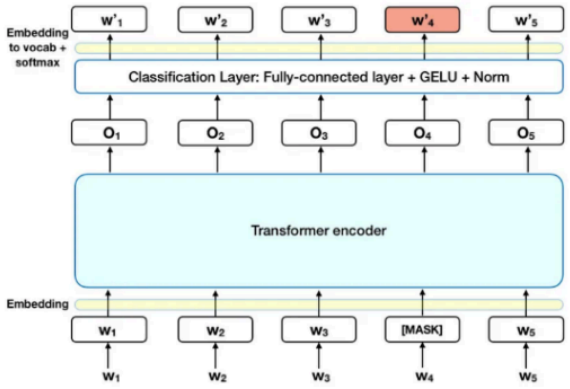
Велики језички модели (енг. *Large language Models – LLM*) су револуционисали комуникацију са системима машинског учења омогућавајући обављање задатака у реалном времену на основу природног језика. Ови модели, који садрже милијарде параметара, тренирани су на огромним количинама текстуалних података, чиме стичу широко разумевање света. За обуку се користи самонадгледано учење (енг. *self-supervised learning*), које омогућава рад са неозначеним подацима и побољшава способност генерисања текста [2]. Након почетког трениранја модели могу решавати задатке користећи *zero-shot* и *few-shot* техника, које им омогућавају да извршавају задатке са мало или без примера. Инжењеринг упита (енг. *prompt engineering*) и техника ланац мисли (енг. *chain of thought – CoT*) омогућавају прецизније одговоре и боље решавање сложенијих проблема. За специфичне задатке у различитим областима, ови модели захтевају додатно дообучавање на доменским подацима.

2.2.4. *BERT*

*BERT (*енг. *Biderectional Encoder Representations from Transformer*) је врста трансформера, где је кључна иновација примена бидирекционог тренирања. Овај приступ тренирања модела у оба смера доводи до резултата који показују да језички модел на овај начин може имати дубље разумевање контекста и тока језика него модел трениран у једном смеру. За разлику од основног модела трансформера, *BERT* користи само кодер. Кодер *BERT* модела обрађује читаву секвенцу речи одједном. Ова карактеристика омогућава моделу да разуме реч у контексту свих околних речи, како с лева, тако и с десна. *BERT* користи две специфичне стратегије тренирања: моделовања маскираног језика и предвиђање наредне реченице [3].

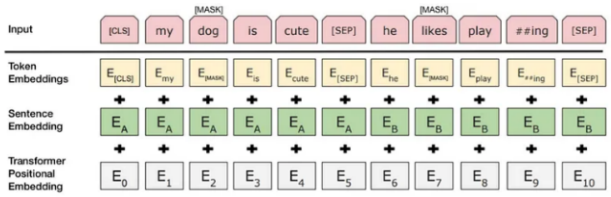
Моделовање маскираног језика подразумева да се 15% речи у низу замени са [MASK] токеном, а модел затим покушава да предвиди те речи на основу контекста. Око 80% замењених речи је означено са [MASK], 10% се замењује случајним речима, а преосталих 10% остаје нетакнуто. Овај приступ помаже моделу да научи контекст, не само предвиђање маскираних речи, тако што га приморава да разуме како се маскиране и немаскиране речи повезују. Предвиђање се постиже додавањем класификационог слоја и применом *softmax* функције како би се израчунала вероватноћа за сваку реч. Ова метода омогућава моделу боље разумевање контекста, иако тренирање може бити

спорије. На слици 3 је приказана описана стратегија тренирања.



Слика 3. *Стратегија тренирања помоћу моделовања маскираног језика*

Стартегија предвиђања следеће реченице у *BERT* моделу користи парове реченица како би модел научио да одреди да ли друга реченица у пару следи прву у оригиналном тексту. У половини случајева, друга реченица је заиста следећа, док у осталих 50% случајева друга реченица је насумично изабрана и обично није повезана с првом. За обраду ових парова, [CLS] токен се додаје на почетак, [SEP] на крај сваке реченице, и сваком токену се додаје ембединг који означава његов положај и реченицу којој припада. Ова стратегија помаже *BERT* моделу да боље разуме контекст и односе између реченица. На слици 4 приказана је обрада улаза за описану стратегију.



Слика 4. *Обрада улаза за стратегију предвиђања следеће реченице*

Да би *BERT* утврдио да ли је друга реченица повезана са првом, цео низ пролази кроз трансформер, при чему се излаз [CLS] токена трансформише у вектор за класификацију. *Softmax* затим израчунава вероватноћу повезаности реченица. Моделовање маскираног језика и предвиђање следеће реченице тренирају се истовремено како би се минимизовала заједничка функција губитка. Овакво удружено тренирање омогућава *BERT* моделу боље разумевање контекста и структуре језика.

Како би *BERT* био успешан при решавању различитих језичких задатака, углавном је довољно додати мали слој на основни модел.

3. КОРИШЋЕНИ СКУПОВИ ПОДАТАКА

За потребу тестирања различитих модела у комуникацији са корисницима, коришћен је скуп података који се састоји од питања и одговора везаних за осигурање. Овај скуп података добијен је уз допуштење клијента. За дотренирање *BERT* модела искоришћен је *Semantic Textual Similarity Benchmark* скуп података [4], који је уједно и један од најчешће коришћених скупова за проналажење сличних питања.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

Најбољи резултати при тестирању модела високог одзива добијени су коришћењем претрениранog *Word2Vec* модела уз сумирање вектора речи и еуклидску удаљеност за мерење сличности.

За модел високе прецизности, чије је унапређење уједно и циљ овог истраживања, спроведено је и квантитативно и квалитативно тестирање перформанси сијамског *LSTM* и *BERT* модела на истом тест скупу који је споменут у претходном поглављу. Квантитативна анализа је показала да *LSTM* често није успевао да обухвати све кључне аспекте корисничког упита, што је доводило до враћања делимично релевантних или потпуно нетачних питања, до је *BERT* постигао значајно боље метрике прецизности и тачности. Квалитативно посматрано, *BERT* је, захваљујући бољем разумевању контекста, доследно проналазио семантички најприближнија питања, што је резултирало знатно адекватнијим и поузданијим одговорима у реалним сценаријима.

5. ЗАКЉУЧАК

Имплементација трансформер модела за четбота у домену осигурања значајно побољшава тачност, брзину и укупно корисничко искуство у односу на претходно имплементирани *LSTM* приступ. Трансформери омогућавају бољу обраду комплексних и дугих упита, што је од великог значаја како у осигурању, тако и у другим областима. У циљу даље применљивости у реалним условима и даљем унапређењу система могуће је интегрисати *RAG* методологију, а такође и применити различите механизме заштите података како би се добило још напредније, флексибилније и безбедније решење којим би се додатно унапредило корисничко искуство.

6. ЛИТЕРАТУРА

[1] Ferrer, J. (2024) How transformers work: A detailed exploration of Transformer architecture, DataCamp.

[2] Nyandwi, J. (2023) Ai Research Blog - The Transformer Blueprint: A holistic guide to the transformer neural network architecture, Deep Learning Revision

[3] Horev, R. (2018) Bert explained: State of the art language model for NLP, Medium.

[4] https://paperswithcode.com/dataset/sts-benchmark

Кратка биографија:

Нађа Кањух рођена је у Врбасу 2000. године. Мастер рад на Факултету техничких наука из области Електротехнике и рачунарства одбранила је 2025. године.

контакт: nadjakanjuh00@gmail.com