UDC: 004.88 Info M str. 11-17

РАЗВОЈ МОДЕЛА ЗА ДЕТЕКЦИЈУ УБРИЗГАВАЊА УПИТА У ЈЕЗИЧКИМ МОДЕЛИМА
DEVELOPING MODELS FOR DETECTING PROMPT INJECTION ATTACKS IN LANGUAGE MODELS

Душица Голубовић

РЕЗИМЕ: Убризгавање упита (енг. prompt injection) представља један од најзначајнијих безбедносних изазова у савременим системима заснованим на великим језичким моделима, јер омогућава злонамерним корисницима да заобиђу ограничења и измене понашање модела. Овај рад испитује могућност примене претходно тренираних језичких модела заснованих на BERT архитектури у задатку детекције убризгавања упита након процеса финог подешавања.
Експериментални део рада обухвата дообуку и евалуацију више BERT-базираних модела на три различита скупа података: јавном скупу преузетом са платформе Kaggle, независном тест скупу са платформе Hugging Face, као и ручно конструисаном скупу упита на српском језику. Перформансе модела процењене су применом стандардних метрика бинарне класификације, са посебним осврtом на одзив и F1 меру због небалансиране расподеле класа и безбедносне природе задатка. Резултати указују на ограничену способност генерализације модела на независном Hugging Face скупу података, док су значајно бољи резултати остварени на српском скупу, што потврђује способност модела да се успешно прилагоде језички специфичном контексту. MiniLM-L6 модел показао је најповољнији однос прецизности и одзива, док је додатна оптимизација хиперпараметара донела само ограничена побољшања.

КЉУЧНЕ РЕЧИ: убризгавање упита, пробијање ограничења, језички модели, трансформер архитектура, BERT модел, фино подешавање, безбедност вештачке интелигенције.

ABSTRACT: Prompt injection attacks represent one of the most critical security challenges in modern systems based on large language models, as they enable malicious users to bypass constraints and manipulate model behavior. This paper investigates the applicability of pretrained BERT-based language models for detecting prompt injection attacks after fine-tuning.

The experimental study includes fine-tuning and evaluation of several transformer-based models on three datasets: a publicly available Kaggle dataset used for training, an independent test dataset obtained from the Hugging Face platform, and a manually constructed dataset of Serbian-language prompts. Model performance was assessed using standard binary classification metrics, with particular emphasis on recall and F1 score due to class imbalance and the security-sensitive nature of the task. The results indicate limited generalization performance on the independent Hugging Face dataset, suggesting sensitivity to dataset distribution and prompt formulation differences. In contrast, significantly better results were achieved on the Serbian dataset, demonstrating the models' ability to adapt effectively to language-specific characteristics after fine-tuning. Among the evaluated models, MiniLM-L6 achieved the most favorable balance between precision and recall, while additional hyperparameter optimization resulted in only marginal improvements. These findings highlight the importance of diverse and representative datasets and the need for further research on improving generalization in prompt injection detection across different languages and domains.

KEY WORDS: prompt injection, jailbreak, language models, transformer architecture, BERT model, fine-tuning, AI security.

1. УВОД

Убризгавање упита представља један од најактуелнијих безбедносних изазова у системима заснованим на језичким моделима, јер омогућава злонамерним корисницима да заобиђу уграђена ограничења и измене очекивано понашање модела. Са све већом применом језичких модела у образовним, комерцијалним и комуникационим системима, питања поузданости, робусности и безбедности постају кључни предуслови њихове практичне употребе.

Проблем убризгавања упита превазилази оквире чисто техничког изазова и представља озбиљан безбедносни ризик, јер може довести до ширења дезинформација, кршења приватности или извршавања нежељених радњи од стране модела [3, 12]. Постојећа истраживања показују да су савремени језички модели осетљиви на пажљиво конструисане уносе који нарушавају њихово понашање, што указује на потребу за систематским методама детекције и евалуације оваквих напада. У литератури се често користи и термин пробијање ограничења (енг. *jailbreak*), којим се означавају упити чији је циљ заобилажење безбедносних механизама и ограничења језичког модела. Иако се појмови убризгавање упита и пробијање ограничења не користе увек као потпуни синоними, у контексту овог рада оба се односе на покушаје манипулације понашањем модела путем пажљиво формулисаних инструкција.

У овом раду анализирана је способност више претходно обучених модела заснованих на BERT архитектури да детектују убризгавање упита након процеса финог подешавања. Посебан акценат стављен је на процену способности генерализације модела на различитим скуповима података, као и на испитивање утицаја хиперпараметара на стабилност и перформансе модела [3].

Експериментална анализа обухвата евалуацију модела на јавним скуповима података на енглеском и немачком језику, као и на ручно конструисаном скупу упита на српском језику. На тај начин, рад пружа увид у применљивост BERT-базираних модела у безбедносном контексту на локалном језику, који је у постојећој литератури недовољно истражен.

2. ТЕОРИЈСКЕ ОСНОВЕ

2.1 *НЕУРОНСКЕ МРЕЖЕ*

Вештачке неуронске мреже представљају основу савремених система машинског учења и инспирисане су начином функционисања биолошких неуронских система [6]. Оне се састоје од великог броја једноставних јединица, неурона, организованих у слојеве и међусобно повезаних тежинама, које се током процеса учења аутоматски прилагођавају на основу улазних података и очекиваних излаза [2].

У контексту обраде природног језика, неуронске мреже омогућавају моделирање сложених, нелинеарних односа између речи и њиховог контекста. [12] Иако су раније архитектуре, као што су рекурентне неуронске мреже, биле широко коришћене за секвенцијалне податке, савремени NLP системи све чешће користе трансформер архитектуру, која показује боље перформансе и стабилност приликом обраде дугих секвенци текста.

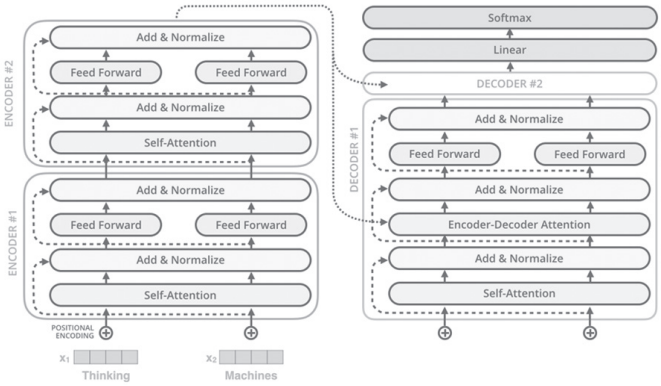
2.2 *ТРАНСФОРМЕР АРХИТЕКТУРА*

Трансформер архитектура представља значајан помак у односу на традиционалне приступе обради секвенцијалних података, јер укида рекурентне и конволутивне слојеве и уводи механизам пажње (енг. *attention*) као кључни елемент модела [13]. Овај механизам омогућава паралелну обраду целе секвенце и ефикасно моделирање дугорочних зависности у тексту, што је од посебног значаја у задацима анализе језика.

Основна идеја механизма самопажње јесте да сваки токен у улазној секвенци добија нову контекстуалну репрезентацију на основу релевантности других токена у истој секвенци [1]. Проширење овог механизма у виду пажње са више глава омогућава моделу да истовремено учи различите типове релација између речи, као што су синтаксички и семантички односи [13].

Трансформер модел се састоји од енкодер и декодер компоненти, при чему се у многим савременим NLP задацима користи искључиво енкодер део архитектуре, што је од посебног значаја у задацима анализе језика. [11]

Слика 1 – Општа структура трансформер архитектуре [1]

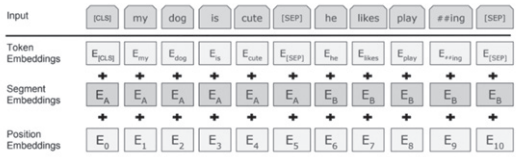


2.3 БЕРТ МОДЕЛ

BERT (енг. *Bidirectional Encoder Representations from Transformers*) представља један од најзначајнијих трансформер модела у области обраде природног језика и заснива се искључиво на енкодер делу трансформер архитектуре [13]. Кључна предност BERT модела лежи у двосмерном моделирању контекста, које омогућава истовремено разматрање речи које претходе и следе посматрани токен, чиме се постиже дубље разумевање семантичких односа у тексту.

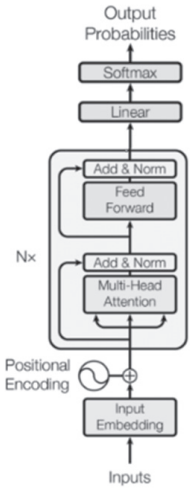
На улазу у модел, текст се представља комбинацијом токенских, позиционих и сегментних уградњи, које заједно формирају улазне ембединг векторе. Овај поступак, приказан на Слици 2, омогућава моделу да, поред значења појединачних речи, узме у обзир и њихов редослед, као и структуру улазне секвенце, што је од кључног значаја за успешно моделирање контекста.

Слика 2 – Процес формирања улазних ембединга у BERT моделу [4]



Захваљујући својој архитектури и могућности ефикасне дообуке, *BERT*-базирани модели су посебно погодни за задатке детекције убризгавања упита, где је неопходно препознати суптилне семантичке и контекстуалне обрасце у тексту. Архитектура BERT модела приказана је на Слици 3.

Слика 3 – Архитектура BERT модела [13]



BERT модел се претходно обучава применом задатака *Masked Language Modeling* и *Next Sentence Prediction*, што му омогућава стицање општег језичког знања. Након тога, модел се може фино подесити за специфичне задатке, као што су класификација текста или детекција убризганих уноса, додавањем једноставног излазног слоја и додатним тренирањем на мањем, циљаном скупу података.

3. ПОСТАВКА ЕКСПЕРИМЕНТА

3.1 ОКРУЖЕЊЕ

Експерименти су изведени у *Google Colab* окружењу, уз коришћење *Tesla T4 GPU* акцелератора, што је омогућило ефикасно тренирање и евалуацију трансформер модела. Имплементација је реализована у програмском језику *Python* 3.10, уз употребу библиотека *Transformers*, *PyTorch*, *Scikit-learn* и *Sentence-Transformers*.

Овакво окружење обезбедило је репродуктивност експеримената, једноставну конфигурацију различитих варијанти модела и пуну интеграцију са *Hugging Face* алатима.

3.2 СКУПОВИ ПОДАТАКА

У истраживању су коришћена три скупа података различитог порекла и језичких карактеристика, са циљем процене способности модела да детектују убризгавање упита и генерализују на податке који нису коришћени током обуке. *Kaggle* скуп података коришћен је за фино подешавање модела, *Hugging Face* скуп за тестирање генерализације на енглеском и немачком језику, док је ручно креирани скуп на српском језику употребљен за испитивање перформанси на локалном језику.

Сви скупови садрже позитивне примере злонамерних упита, и неутралне дијалошке уносе, са различитим пропорцијама и дужинама текстова.

Подаци су припремљени за бинарну класификацију, при чему су примери убризгавања упита означени као позитивна класа (1), а остали као негативна (0). Скупови су учитани у *pandas DataFrame*, очишћени од недостајућих вредности и насумично измешани. Скуп за обуку је стратификовано подељен на тренинг и валидациони подскуп, након чега су подаци конвертовани у *DatasetDict* формат. Текстови су токенизовани уз примену параметара *truncation=True*, *padding=True* и *max_length=128*.

Табела 1 – *Приказ скупова података*

Порекло скупа података	Укупан број примера	Примери убризгавања упита (%)	Језик
Kaggle	1581	12,6	Енглески
Hugging Face	546	37,2	Енглески и немачки
Ручно креиран	85	42,4	Српски

Ручно конструисани српски скуп података садржи укупно 85 примера, од чега је 42,4% означено као позитивна класа. Примери су ручно креирани на основу типичних образаца убризгавања упита из енглеског скупа података, уз њихову адаптацију српском језичком и културолошком контексту. Упити обухватају покушаје заобилажења ограничења модела, генерисања забрањеног садржаја и модификације системских инструкција. Иако овакав приступ омогућава контролисану евалуацију модела на локалном језику, могуће је присуство пристрасности услед ограничене величине скупа и ручног начина конструкције примера.

Иако сва три скупа података садрже позитивне и негативне примере, међу њима постоје значајне разлике у начину комуникације и структури упита. *Kaggle* скуп доминантно садржи експлицитне покушаје убризгавања упита, често засноване на директним инструкцијама, попут *„Ignore previous instructions”,* или захтевима за активацију алтернативних режима рада модела. *Hugging Face* скуп обухвата разноврсније и имплицитније примере, укључујући сценарије преузимања улога, промену идентитета модела, политичке и идеолошке теме, као и посредне покушаје заобилажења ограничења. Ручно креирани српски скуп креиран је по узору на обрасце присутне у *Kaggle* скупу, уз прилагођавање језичком и културном контексту.

Табела 2 – *Упоредни преглед карактеристика коришћених скупова података*

Карактеристика	Kaggle	Hugging Face	Српски
Стил напада	Експлицитан, директан	Имплицитан, концептуалан	Експлицитан (по узору на *Kaggle*)
Анотација	Јавна	Јавна	Ручна
Језичка разноликост	ниска	висока	Ниска

3.3 МОДЕЛИ

У истраживању су коришћени претходно тренирани трансформер модели различитих величина и архитектура, како би се упоредиле њихове перформансе и рачунарски захтеви у задатку детекције убризгавања упита. У анализу су укључена четири модела: два *MiniLM* модела из *SentenceTransformers* библиотеке (*all-MiniLM-L6-v2* са приближно 22 милиона параметара [8] и *paraphrase-multilingual-MiniLM-L12-v2* са око 118 милиона параметара [9]), *GIST-small* модел са око 33 милиона параметара [10] оптимизован за контрастивно учење, као и *bert-base-uncased-eurlex* заснован на BERT-BASE архитектури са приближно 110 милиона параметара [7]. Иако припадају различитим имплементационим варијантама трансформер модела, сви модели се заснивају на енкодер архитектури и могу се посматрати као *BERT*-деривати у ширем смислу.

Сви модели су дообучени коришћењем *Trainer* класе, уз исте параметре токенизације (*truncation=True*, *padding=True*, *max_length=128*). Тренинг је изведен у трајању од три епохе, са величином серије 8, стопом учења 5e-5 и применом механизма раног заустављања.

Табела 3 – *Почетна конфигурација хиперпараметара за фино подешавање модела*

Хиперпараметар	Вредност
Стопа учења (енг. *learning rate*)	5e-5
Величина серије (енг. *batch_size*)	8
Број епоха (енг. *num_train_epochs*)	3
Рано заустављање тренинга (енг. *early_stopping_patience*)	2
Максимална дужина секвенце (енг. max_length)	128
Слабљење тежина (енг. *weight_decay*)	0.01

Евалуација је спроведена применом стандардних метрика бинарне класификације (тачност, прецизност, одзив, F1 и AUC) на валидационом делу тренинг скупа и на два независна тестна скупа. Посебан значај дат је одзиву и F1 метрици, имајући у виду небалансирану расподелу класа и безбедносну природу задатка.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

4.1. *ОСНОВНИ ЕКСПЕРИМЕНТ: ПЕРФОРМАНСЕ МОДЕЛА СА ПОДРАЗУМЕВАНИМ ХИПЕРПАРАМЕТРИМА*

Циљ основног експеримента био је процена перформанси четири претходно тренирана трансформер модела након финог подешавања на *Kaggle* скупу података, као и идентификација модела који постиже најбољи баланс између перформанси и рачунарске ефикасности. Сви модели су обучавани под истим условима, уз подразумеване хиперпараметре (величина серије 8, стопа учења 5e-5, 3 епохе, максимална дужина секвенце 128) и стратификовану поделу на *train* и *validation* подскупове.

Валидационе перформансе и калсификационе метрике модела на *Kaggle* скупу приказани су у Табели 4 и у Табели 5.

Табела 4 – *Валидационе перформансе модела*

Модел	Просечан губитак на валидационом скупу	Број обрађених узорака у секунди током евалуације	Најбољи checkpoint
sentence-transformer/all-MiniLM-L6-v2	0,007	709,239	Checkpoint-300
avsolatorio/GIST-small-Embedding-v0	0,029	206,094	Checkpoint-199
nlpaueb/bert-base-uncased-eurlex	0,016	127,293	Checkpoint-300
sentence-transformer/paraphrase-multilingual-MiniLM-L12-v2	0,003	366,042	Checkpoint-200

Табела 5 - *Класификационе метрике модела*

Модел	Тачност	Прецизност	Одзив	F1 мера	AUC
sentence-transformer/all-MiniLM-L6-v2	0,997	1	0,973	0,984	1
avsolatorio/GIST-small-Embedding-v0	0,997	1	0,973	0,986	0,998
nlpaueb/bert-base-uncased-eurlex	0,997	1	0,973	0,986	0,999
sentence-transformer/paraphrase-multilingual-MiniLM-L12-v2	1	1	1	1	1

Сви модели постижу изузетно високе вредности тачности, F1 мере и AUC метрике, што указује на њихову способност да ефикасно науче обрасце присутне у тренинг подацима. Истовремено, уочљиве су разлике у рачунарској ефикасности, при чему *MiniLM* варијанте постижу значајно већу брзину евалуације у поређењу са већим моделима, што их чини погоднијим за примене са ограниченим ресурсима.

Иако су овакви резултати технички пожељни, веома ниске вредности функције губитка и готово савршене метрике могу указивати на претерано прилагођавање моделâ подацима на којима су обучавани, што је чест проблем у практичним применама машинског учења [5]. Због тога је било неопходно спровести додатну евалуацију на независним тест скуповима података, како би се поуздано проценила способност генерализације.

4.1.1 *Резултати на Hugging Face скупу података*

Резултати евалуације на *Hugging Face* скупу података приказани су у Табели 6.

Табела 6 – *Тест евалуационе метрике на Hugging Face скупу података*

Модел	Тачност	Прецизност	Одзив	F1 мера	AUC
sentence-transformer/all-MiniLM-L6-v2	0,689	0,837	0,201	0,325	0,611
avsolatorio/GIST-small-Embedding-v0	0,648	0,923	0,059	0,111	0,823
nlpaueb/bert-base-uncased-eurlex	0,688	0,837	0,202	0,325	0,611
sentence-transformer/paraphrase-multilingual-MiniLM-L12-v2	0,671	0,925	0,123	0,217	0,667

За разлику од резултата на тренинг скупу, овде се уочава значајан пад перформанси код свих модела, што указује на ограничену способност генерализације на податке из другог домена. Док су на валидационом подскупу *Kaggle* скупа сви модели остварили F1 вредности између 0,984 и 1 (Табела 5), на Hugging Face скупу оне опадају на интервал од 0,111 до 0,325 (Табела 6). Овај пад може указивати на могуће претерано прилагођавање моделâ обрасцима присутним у тренинг подацима, као и на осетљивост на промену дистрибуције и структуре улазних упита.

Иако сви модели постижу сличну тачност (у распону од 0,65 до 0,69), анализа појединачних метрика открива значајне разлике у њиховом понашању. Модели *bert-base-uncased-eurlex, GIST-small* и *MiniLM-L12* показују високу прецизност, али изузетно низак одзив, што указује на конзервативну стратегију класификације. Они ретко означавају пример као убризгавање упита, али када то учине, класификација је најчешће тачна. Последично, њихове F1 вредности остају ниске.

Насупрот томе, *MiniLM-L6* модел постиже највиши одзив и највећу F1 меру међу тестираним моделима, иако уз нешто нижу прецизност. Овај компромис резултира бољим укупним балансом перформанси и указује на већу способност препознавања позитивних примера у непознатом домену. У целини, резултати показују да модели обучени на *Kaggle* скупу нису у потпуности генерализовали на *Hugging Face* податке, што се може приписати разликама у стилу, садржају и формализацији убризганих упита.

Разлика у перформансама може се објаснити и разликама у структури скупова података. Док Kaggle скуп доминантно садржи експлицитне обрасце убризгавања упита, Hugging Face скуп укључује већи број имплицитних и краћих, али контекстуално сложенијих примера. Такви примери отежавају детекцију и представљају изазов за генераизацију модела.

4.1.2 *Резултати на ручно креираном српском скупу података*

Резултати евалуације на ручно креираном српском скупу података су приказани у Табели 7, показују значајно боље перформансе свих модела у поређењу са *Hugging Face* скупом. Овај налаз је посебно значајан имајући у виду да ниједан од модела није био експлицитно обучаван на српском језику.

Табела 7 – *Тест евалуационе метрике на српском скупу података*

Модел	Тачност	Прецизност	Одзив	F1 мера	AUC
sentence-transformer/all-MiniLM-L6-v2	0,953	1	0,889	0,941	0,975
avsolatorio/GIST-small-Embedding-v0	0,847	1	0,639	0,78	0,956
nlpaueb/bert-base-uncased-eurlex	0,8	1	0,528	0,69	0,983
sentence-transformer/paraphrase-multilingual-MiniLM-L12-v2	0,941	0,969	0,889	0,927	0,982

MiniLM модели постижу највише вредности свих кључних метрика, при чему *all-MiniLM-L6-v2* остварује најбољи однос тачности, одзива и F1 мере. Вишејезички *MiniLM-L12* модел такође показује високе перформансе, што је у складу са очекивањима о његовој способности језичке генерализације. Модели *GIST-small* и *bert-base-uncased-eurlex* такође бележе значајан раст F1 и AUC вредности у односу на *Hugging Face* резултате.

Резултати указују да убризгани упити поседују структурне и семантичке обрасце који нису у потпуности везани за конкретан језик. Због тога модели могу пренети део наученог знања између различитих језичких контекста. Истовремено, високе вредности метрика могу бити делимично последица ограничене величине и јасне структуре српског скупа података, што представља важно ограничење које треба узети у обзир при тумачењу резултата. Такође, чињеница да модели остварују високе резултате и на скупу података који није коришћен током обуке указује да пад перформанси на *Hugging Face* скупу није нужно последица искључиво претераног прилагођавања, већ и поменутих значајних разлика у начину формулације и комплексности убризганих упита.

4.2 *ХИПЕРПАРАМЕТАРСКА ОПТИМИЗАЦИЈА НАЈУСПЕШНИЈИХ МОДЕЛА*

Након анализе основних експеримената са подразумеваним вредностима хиперпараметара (поглавље 4.1), у овом делу рада испитан је утицај промене конфигурације тренинга на перформансе два најуспешнија модела – *all-MiniLM-L6-v2* и *paraphrase-multilingual-MiniLM-L12-v2*. Ови модели су одабрани на основу повољног односа перформанси и рачунарске ефикасности у претходним експериментима.

Циљ хиперпараметарске оптимизације није био искључиво постизање виших вредности евалуационих метрика, већ и разумевање начина на који поједини параметри тренинга утичу на понашање модела у задацима детекције убризгавања упита, посебно у контексту генерализације на податке из другог домена.

Оптимизација је спроведена применом мрежне претраге (енг. *grid search*), при чему су систематски вариране вредности стопе учења, величине серије, броја епоха и максималне дужине улазних секвенци, док су остали параметри задржани из основне конфигурације. Оваквим приступом омогућено је поређење различитих комбинација хиперпараметара под истим експерименталним условима и анализа њиховог утицаја на кључне метрике перформанси.

4.2.1 *Перформансе L6 модела на Hugging Face скупу података*

У овом експерименту анализиране су перформансе модела *all-MiniLM-L6-v2* на *Hugging Face* скупу података, са циљем испитивања утицаја различитих конфигурација хиперпараметара на способност генерализације. Резултати су приказани у Табели 8.

Табела 8 – *Перформансе различитих конфигурација MiniLM-L6 модела на Hugging Face скупу података*

Конфигурација *MiniLM-L6* модела	Тачност	Прецизност	Одзив	F1 мера	AUC
lr 1e-4, bs 8, ep 3	0,694	0,86	0,212	0,340	0,673
lr 2e-5, bs 8, ep 3	0,679	0,938	0,148	0,255	0,588
lr 5e-5, bs 16, ep 3	0,7	0,767	0,276	0,406	0,637
lr 5e-5, bs 16, ga 2, ep 3	0,665	0,955	0,103	0,187	0,602
lr 5e-5, bs 8, ep 10 ★	**0,718**	**0,915**	**0,266**	**0,412**	**0,701**
lr 5e-5, bs 8, ep 3	0,7	0,882	0,222	0,354	0,641
lr 5e-5, bs 8, ep 3, ml 256	0,658	0,944	0,084	0,154	0,624
lr 5e-5, bs 8, ep 6	0,709	0,907	0,241	0,381	0,665
lr 5e-6, bs 8, ep 3	0,636	1	0,02	0,039	0,595

Све тестиране конфигурације показују умерене перформансе, уз ограничен одзив и релативно високу прецизност. Најбољи укупни резултат постигнут је конфигурацијом са стопом учења 5e-5, величином серије 8 и 10 епоха тренинга, при чему су остварене вредности F1 мере од 0,412 и AUC од 0,701.

У поређењу са резултатима пре оптимизације, уочава се побољшање у готово свим метрикама, при чему је најизраженији напредак забележен код F1 и AUC вредности. Ово указује на бољу равнотежу између прецизности и одзива након подешавања параметара, што је посебно значајно у контексту детекције ретких и критичних случајева убризгавања упита.

Ипак, висока прецизност уз релативно низак одзив указује на конзервативно понашање модела, који избегава означавање упита као убризганих уколико није високо сигуран у своју процену. Повећање броја епоха показало се као најефикаснији механизам побољшања перформанси, док су веће величине серије и градијентна акумулација имале ограничен или негативан утицај.

4.2.2 *Перформансе L6 модела на ручно креираном српском скупу података*

Перформансе *L6* модела на ручно креираном српском скупу података приказане су у Табели 9. За разлику од резултата добијених на *Hugging Face* скупу, овде се уочава значајно побољшање свих евалуационих метрика.

Табела 9 – *Перформансе различитих конфигурација L6 модела на српском скупу података*

Конфигурација *MiniLM-L6* модела	Тачност	Прецизност	Одзив	F1 мера	AUC
lr 1e-4, bs 8, ep 3	0,906	0,967	0,806	0,879	0,976
lr 2e-5, bs 8, ep 3	0,929	1	0,833	0,909	0,972
lr 5e-5, bs 16, ep 3	0,918	0,968	0,833	0,896	0,978
lr 5e-5, bs 16, ga 2, ep 3	0,894	1	0,75	0,857	0,975
lr 5e-5, bs 8, ep 10 ★	**0,941**	**0,97**	**0,889**	**0,928**	**0,973**
lr 5e-5, bs 8, ep 3	0,929	1	0,833	0,909	0,975
lr 5e-5, bs 8, ep 3, ml 256	0,812	1	0,556	0,714	0,973
lr 5e-5, bs 8, ep 6	0,929	0,969	0,861	0,912	0,977
lr 5e-6, bs 8, ep 3	0,635	1	0,139	0,244	0,972

Најбоља конфигурација (стопа учења 5e-5, величина серије 8, 10 епоха) остварила је F1 меру од 0,928 и AUC од 0,973, што указује на висок ниво стабилности и поузданости модела. Високе вредности прецизности и одзива показују да модел успешно идентификује убризгане упите, уз минималан број лажно позитивних примера.

Поређење са резултатима на *Hugging Face* скупу показује да модел знатно боље функционише на подацима који су језички и структурно конзистентнији. Ово указује на то да да одређени шаблони убризгавања могу бити делимично преносиви између језичких домена, али да перформансе значајно зависе од дистрибуције и формализације података.

4.2.3 *Перформансе L12 модела на Hugging Face скупу података*

Резултати оптимизације хиперпараметара за модел *paraphrase-multilingual-MiniLM-L12-v2* приказани су у Табели 10. Иако је реч о дубљој и комплекснијој архитектури, постигнуте перформансе остају ограничене.

Табела 10 – *Перформансе различитих конфигурација L12 модела на Hugging Face скупу података*

Конфигурација *MiniLM-L12* модела	Тачност	Прецизност	Одзив	F1 мера	AUC
lr 1e-4, bs 8, ep 3	0,674	0,931	0,133	0,233	0,799
lr 2e-5, bs 8, ep 3	0,679	0,85	0,167	0,28	0,581
lr 5e-5, bs 16, ep 3	0,667	0,839	0,128	0,222	0,566
lr 5e-5, bs 16, ga 2, ep 3	0,663	0,852	0,113	0,2	0,566
lr 5e-5, bs 8, ep 10	0,676	0,861	0,153	0,259	0,615
lr 5e-5, bs 8, ep 3	0,672	0,816	0,153	0,257	0,623
lr 5e-5, bs 8, ep 3, ml 256 ★	**0,696**	**0,814**	**0,236**	**0,366**	**0,65**
lr 5e-5, bs 8, ep 6	0,661	0,821	0,113	0,199	0,573
lr 5e-6, bs 8, ep 3	0,643	0,786	0,054	0,101	0,482

Најбоља конфигурација остварила је F1 меру од 0,366 и AUC од 0,650, што представља умерено побољшање у односу на основну конфигурацију, пре свега кроз повећање одзива. Ипак, укупне перформансе *L12* модела не премашују резултате постигнуте *L6* варијантом.

Ови налази указују да повећање дубине модела не гарантује бољу дискриминативну способност у задацима детекције убризгавања упита. У конкретном случају, мањи и ефикаснији *L6* модел показује повољнији однос између прецизности, одзива и рачунарских захтева.

Укупно посматрано, резултати хиперпараметарске оптимизације показују да повећање дубине модела не доводи нужно до побољшања перформанси у задацима детекције убризгавања упита. Иако *L12* модел поседује већу архитектурну сложеност, у више конфигурација мањи *L6* модел постиже бољи баланс између прецизности и одзива, као и више F1 вредности.

Слични трендови уочени су и приликом евалуације на ручно креираном српском скупу података, где *L6* модел показује високу стабилност и поузданост. Ови налази указују да ефикасност модела у овом домену више зависи од структуре података и карактеристика задатка него од саме дубине архитектуре.

5. ЗАКЉУЧАК

Спроведено истраживање показало је да су језички модели способни да у извесној мери препознају убризгане упите. Међутим, њихова ефикасност у великој мери зависи од при-

роде и дистрибуције података на којима су тренирани. Док су резултати на *Hugging Face* скупу указали на ограничену способност генерализације услед разлика у дефиницијама и облицима убризганих упита, перформансе на ручно креираном српском скупу података биле су стабилније и показале бољу усклађеност између прецизности и одзива.

Модел *MiniLM-L6* се издвојио као најуспешнији модел у погледу укупног баланса евалуационих метрика, што потврђује његову способност да поуздано препозна различите типове убризганих инструкција уз задржавање релативно ниске стопе погрешних класификација. Хиперпараметарска оптимизација допринела је додатном унапређењу резултата, али је истовремено показала да сама техничка подешавања нису довољна за постизање значајно веће отпорности модела.

Добијени резултати указују да би даље истраживање требало да буде усмерено ка развоју разноврснијих и репрезентативнијих скупова података, који обухватају различите језичке форме, стилове и нивое комплексности убризгавања упита. Поред тога, перспективу представља примена напреднијих приступа, као што су контекстуално учење (енг. *in-context learning*), дообука на специфичним доменима, као и интеграција више модела у јединствен систем (енг. *ensemble learning*), ради постизања веће поузданости и робусности детекције убризганих упита.

Ови закључци потврђују значај теме у контексту безбедности великих језичких модела и наглашавају потребу за њиховим континуираним тестирањем и унапређивањем у циљу изградње поузданијих и отпорнијих система за обраду природног језика.

6. ЛИТЕРАТУРА

[1] Alammar, J. (2018). *The Illustrated Transformer*. Преузето са https://jalammar.github.io/illustrated-transformer/ [датум приступа: 18.8.2025.]

[2] Bishop, C. M., & Bishop, H. (2023). *Deep Learning: Foundations and Concepts*. Springer Nature.

[3] Blevins, T., & Zettlemoyer, L. (2022). Language contamination helps explain the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3563–3574). Association for Computational Linguistics.

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[5] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

[6] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[7] Hugging Face. (2019). *nlpaueb/bert-base-uncased-eurlex* [Model]. Hugging Face Hub. https://huggingface.co/nlpaueb/bert-base-uncased-eurlex [датум приступа: 9.9.2025.]

[8] Hugging Face. (2022). *sentence-transformers/all-MiniLM-L6-v2* [Model]. Hugging Face Hub. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 [датум приступа: 12.9.2025.]

[9] Hugging Face. (2022). *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* [Model]. Hugging Face Hub. https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [датум приступа: 12.9.2025.]

[10] Hugging Face. (2024). *avsolatorio/GIST-small-Embedding-v0* [Model]. Hugging Face Hub. https://huggingface.co/avsolatorio/GIST-small-Embedding-v0 [датум приступа: 10.9.2025.]

[11] Khalid, S. (2020). Applications of transformer models in natural language processing. *Journal of Artificial Intelligence Research*, 69, 1–21. https://doi.org/10.1613/jair.1.12345 [датум приступа: 27.8.2025.]

[12] Nielsen, M. (2015). *Neural Networks and Deep Learning*.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (NeurIPS 2017).

Душица Голубовић, Yettel d.o.o.
Контакт: dusicabgolubovic@gmail.com
Област интересовања: наука о подацима, машинско учење, обрада природног језика