



Príloha A: Správa o dosiahnutých merateľných ukazovateľoch

Názov projektu	Robustnosť indikátorov dezinformačného obsahu generovaného AI vo viacjazyčnom online priestore
Akronym	RobIndAI
Kód projektu	09I01-03-V04-00059
Začiatok projektu	01. 11. 2024
Trvanie projektu	20 mesiacov



**Financované
Európskou úniou**
NextGenerationEU

PLÁN [OBNOVY]

Obsah

1	Ciele.....	3
2	Výstupy	5
3	Míľníky.....	6
4	Monitorované údaje dopadu	7
5	Kľúčové ukazovatele komunikácie, diseminácie a exploitácie	9
4	Referencie.....	10

1 Ciele

Cieľom projektu RobIndAI bol výskum metód a modelov umelej inteligencie na zvýšenie robustnosti indikátorov dezinformačného obsahu (z webu a sociálnych médií) s orientáciou najmä na detekciu strojovo-generovaného textu v stredoeurópskych jazykoch. Tento všeobecný cieľ bol rozdelený do troch hlavných cieľov projektu RobIndAI, zosumarizovaných v Tab. 1.

Cieľ 1: Detekcia strojovo generovaného textu		Výsledok
Popis:	Metódy detekcie strojovo generovaného textu sú prioritne zamerané na anglický jazyk (kvôli dostupnosti textov a podpore nástrojov na spracovanie prirodzeného jazyka). V projekte VIGILANT je v tejto oblasti snaha priniesť multilingválne riešenie, ktoré je však zamerané na jazyky policajných zložiek zapojených do projektu. Nutnou podmienkou vyhodnotenia existujúcich detekčných metód v iných jazykoch je dostupnosť označovaných dát. V rámci RobIndAI je preto potrebné vytvoriť vhodný dataset s dostatočným množstvom textov v stredoeurópskych jazykoch a s lokálne relevantným obsahovým kontextom. Následne je potrebné realizovať porovnávaciu štúdiu existujúcich detekčných metód použitím vytvoreného datasetu a natrénovať špecializované detekčné modely na tréningových dátach v daných jazykoch.	splnené
Očakávané výstupy:	Dataset obsahujúci človekom písané texty a strojovo generované texty v stredoeurópskych jazykoch. AI model schopný detegovať strojovo generovaný text v týchto jazykoch. V rámci výskumu budeme kladť dôraz na rozvoj pokročilých jazykových modelov na základe najnovších technológií (napr. GPT-4, Llama2).	splnené
Cieľ 2: Zvýšenie robustnosti detekčných metód		Výsledok
Popis:	Na zvýšenie robustnosti existujúcich detekčných metód je potrebné najskôr vyhodnotenie ich základnej odolnosti voči technikám zabránenia detekcie (napr. homoglyfné útoky, parafrázovanie). Zvýšenie základnej odolnosti bude dosiahnuté doplnením tréningovej časti datasetu o útočné vzorky v stredoeurópskych jazykoch.	splnené
Očakávané výstupy:	Rozšírený dataset o útočné vzorky. Vyhodnotenie odolnosti AI modelov detekcie strojovo generovaných textov v stredoeurópskych jazykoch voči existujúcim útokom a technikám zabránenia detekcie. AI model so zvýšenou robustnosťou, natrénovaný aj na útočných vzorkách v stredoeurópskych jazykoch.	splnené
Cieľ 3: Výber vhodnej architektúry		Výsledok
Popis:	Posledná fáza projektu RobIndAI bude zameraná na výskum detekcie strojovo generovaných textov v stredoeurópskych jazykoch z hľadiska architektúry riešenia. Porovnané budú architektúry s jedným multilingválnym AI modelom a architektúry s jednojazyčným AI modelom dedikovaným pre každý jazyk. Porovnanie bude zamerané na celkovú úspešnosť detekcie pre všetky jazyky zahrnuté do vytvoreného datasetu.	splnené
Očakávané výstupy:	Optimálna architektúra detektora strojovo generovaného textu v stredoeurópskych jazykoch.	splnené

Tab. 1 Prehľad cieľov projektu.

Cieľ 1 bol splnený prostredníctvom realizácie porovnávacej štúdie CEAID [1] (V3.1) a vytvorenia zodpovedajúceho datasetu pokrývajúceho textovú doménu novinových článkov a doménu textov sociálnych médií, pokrývajúcich 7 jazykov stredoeurópskeho regiónu: češtinu, chorvátčinu, maďarčinu, nemčinu, poľštinu, slovenčinu a slovinčinu (V2.1). Štúdia priniesla 5 špecializovaných AI detektorov dotrénovaných na vzorkách textu dedikovaných pre konkrétny jazyk, ako aj množstvo detektorov dotrénovaných na kombináciách týchto piatich trénovacích jazykov. Vzhľadom na veľké množstvo detektorov, tieto boli zverejnené prostredníctvom zverejnenia trénovacích dát (CEAID dataset na platforme Zenodo) a trénovacích zdrojových kódov umožňujúcich plnú replikáciu modelov (GitHub), odkazy dostupné v správe V2.1.

Cieľ 2 bol splnený prostredníctvom návrhu škálovateľnej metodiky trénovania robustných viacjazyčných detektorov založených na jazykových modeloch [2], ako aj prostredníctvom rozšírenia datasetu CEAID o útočné vzorky pomocou homoglyfného útoku a parafrázovania [1] (V3.1). Takáto augmentovaná časť datasetu obsahuje vyváženú vzorku 100 textov pre každú triedu (strojovo-generované texty a ľudské texty), doménu a jazyk z pôvodného CEAID datasetu, ktorá je doplnená o rovnaký počet vzoriek modifikovaných pomocou homoglyfného útoku a pomocou parafrázovacieho útoku. Obidva datasety, pôvodný CEAID aj jeho rošienie o útočné vzorky boli publikované v anonymizovanej forme na platforme Zenodo. Najlepší model vyplývajúci zo štúdie predstavenia metodiky trénovania robustných viacjazyčných detektorov je publikovaný na platforme HuggingFace. Všetky tieto výskumné artefakty sú zdokumentované v správe V2.1 spolu s príslušnými odkazmi.

Cieľ 3 bol splnený prostredníctvom realizácie porovnávacej štúdie zameranej na porovnanie troch rôznych možností realizácie viacjazyčného detekčného systému (V3.2), ktorá identifikovala optimálnu architektúru pre konkrétnu úlohu binárnej detekcie strojovo-generovaného textu v stredoeurópskych jazykoch s ohľadom na úspešnosť detekcie, ako aj výpočtovú náročnosť pri nasadení takéhoto systému do produkcie. Očakávaný výstup je dosiahnutý vo forme odporúčania vzhľadom na požiadavky (resp. obmedzenia) vo výskumnej správe.

2 Výstupy

V Tab. 2 je zosumarizovaný prehľad výstupov projektu spolu s výsledkom ich splnenia.

Číslo výstupu	Výstup	Kód pracovného balíka	Dodanie výstupu	Výsledok
V1.1	Priebežná správa o implementácii a dosiahnutých výsledkoch projektu	NV1	M12	splnené
V1.2	Záverečná správa o dosiahnutých výsledkoch projektu	NV1	M20	splnené
V2.1	Správa o výsledkoch komunikácie, diseminácie a exploitácie	NV2	M20	splnené
V3.1	Výskumná správa o modeloch a metódach robustnej detekcie strojovo-generovaného textu	NV3	M16	splnené
V3.2	Výskumná správa o optimalizovanej architektúre systému na detekciu strojovo-generovaného textu	NV3	M20	splnené

Tab. 1 Prehľad výstupov projektu.

Priebežná správa o implementácii a dosiahnutých výsledkoch projektu (V1.1) bola nahradená oficiálnou priebežnou monitorovacou správou v predpísanom formulári (podľa oficiálneho pokynu). Podobne aj Záverečná správa o dosiahnutých výsledkoch projektu (V1.2) bola nahradená záverečnou monitorovacou správou. Obidve monitorovacie správy, ako aj ostatné výstupy (predstavujúce samostatné dokumenty), boli zverejnené na [webstránke projektu](#) v mesiaci riešenia projektu podľa plánu dodania výstupu, a teda všetky výstupy môžeme pokladať za **splnené**.

3 Mílniky

V Tab. 3 je zosumarizovaný prehľad mílnikov projektu spolu s výsledkom ich splnenia.

Číslo mílnika	Mílnik	Kód pracovného balíka	Spôsob overenia	Dosiahnutie mílnika	Výsledok
ML1	Prvé verzie metód a modelov	NV1	V1.1	M12	splnené
ML2	Hotové metódy detekcie strojovo-generovaných textov	NV3	V3.1	M16	splnené
ML3	Hotová optimalizovaná architektúra detekčného systému	NV3	V3.2	M20	splnené
ML4	Ukončenie projektu	NV1, NV2	V1.2, V2.1	M20	splnené

Tab. 3 Prehľad mílnikov projektu.

S ohľadom na včasné dodanie zodpovedajúcich výstupov (viď Kap. 2) boli všetky mílniky **splnené** v plnej miere podľa plánu.

4 Monitorované údaje dopadu

V Tab. 4 sú prehľadne uvedené očakávané výsledky a dopady spolu s monitorovanými údajmi a zodpovedajúci výsledok či dané ciele boli splnené.

Očakávané výsledky a dopady	Monitorované údaje	Výsledok
Metódy a modely detekcie strojovo-generovaných textov	Zvýšenie presnosti detekcie strojovo-generovaných textov nad úroveň 80% pri stredoeurópsky relevantných dátových vzorkách. Zvýšenie odolnosti voči realizovaným útokom o 50-100%.	splnené
Zverejnenie relevantných dát a výsledkov v rámci otvorenej vedy	Zdieľanie 2 datasetov (1x dataset obsahujúci človekom písané texty a strojovo generované texty v stredoeurópskych jazykoch, 1x dataset doplnený o útočné vzorky v stredoeurópskych jazykoch), ≥30 referencií na výsledky projektu v iných vedeckých prácach. Publikácia 3 nových článkov v oblasti detekcie strojovo-generovaných textov v dôsledku výskumu nadväzujúceho na naše výsledky.	splnené
Spolupráca s novinármi, mediálnymi organizáciami a vedeckými odborníkmi na Slovensku a v zahraničí	Zvýšenie počtu nových spoluprác na domácich aj zahraničných vedeckých projektoch (najmenej 1 podaný Horizon Europe projekt nadväzujúci na RobIndAI). Zvýšená spolupráca s novinármi, overovateľmi faktov a mediálnymi organizáciami v rámci Slovenska, i v zahraničí.	splnené
Zlepšenie aktuálneho stavu poznania (state-of-the-art) v oblasti detekcie strojovo-generovaného textu v stredoeurópskych jazykoch	Zvýšenie úspešnosti detekčných modelov na jazykoch stredoeurópskeho priestoru o 5-10% v porovnaní s už existujúcimi riešeniami v oblasti automatizovanej detekcie strojovo-generovaných textov.	splnené

Tab. 4 Prehľad monitorovaných údajov dopadu.

Zvýšenie presnosti detekcie strojovo-generovaných textov nad úroveň 80 % pri stredoeurópsky relevantných dátových vzorkách bolo dosiahnuté v realizovanej štúdii CEAIID [1] (V3.1, Obr. 1) pri hodnotách metriky AUC ROC našich dotrénovaných detektorov, kde pri tých najlepších dosahovala hodnoty vysoko nad úroveň 90 %. Zvýšenie odolnosti voči realizovaným útokom o 50-100 % je možné pozorovať v rovnakej štúdii (V3.1, Obr. 3) pri porovnaní najlepšieho existujúceho (state-of-the-art) detektora Fast-DetectGPT a nášho najlepšieho dotrénovaného detektora Llama-3.2-3B (de-hu-pl) z hľadiska zníženia úspešnosti detekcie pri modifikácii testovacej množiny pomocou útočných metód. V priemere sa úspešnosť detekcie Fast-DetectGPT znížila o 50,7 %, zatiaľ čo úspešnosť detekcie nášho detektora Llama-3.2-3B (de-hu-pl) sa znížila v priemere o 1,4 %, čo predstavuje relatívne zvýšenie odolnosti o 97,2 %.

Zdieľanie 2 datasetov (1x dataset obsahujúci človekom písané texty a strojovo generované texty v stredoeurópskych jazykoch, 1x dataset doplnený o útočné vzorky v stredoeurópskych jazykoch) je preukázané v správe V2.1 (Tab. 9), kde reportujeme, že oba datasety sú úspešne publikované na platforme Zenodo v anonymizovanej forme na výskumné účely. Údaj ≥ 30 referencií na výsledky projektu v iných vedeckých prácach bude odmeraný až dva roky po skončení projektu (ako bolo stanovené v diseminačných KPI, V2.1), už v súčasnosti je však z polovice naplnený, takže splnenie tohto plánu je vysoko pravdepodobné. Publikácia 3 nových článkov v oblasti detekcie strojovo-generovaných textov v dôsledku výskumu nadväzujúceho na naše výsledky je plne naplnená viacerými štúdiami realizovanými v projekte RobIndAI a publikovaných vo forme vedeckých článkov (viď Kap. 5).

Zvýšenie počtu nových spoluprác na domácich aj zahraničných vedeckých projektoch (najmenej 1 podaný Horizon Europe projekt nadväzujúci na RobIndAI) je splnené podaním návrhu projektu Horizon Europe FIMI-RESIST, ako aj podaním nového návrhu európskeho grantového projektu v schéme Marie Curie Sklodowska Actions. Zvýšená spolupráca s novinármi, overovateľmi faktov a mediálnymi organizáciami v rámci Slovenska, i v zahraničí je preukázaná množstvom a charakterom účastí na odborných fórach a okrúhlych stoloch reportovaných v správe V2.1 (Tab. 11).

Zvýšenie úspešnosti detekčných modelov na jazykoch stredoeurópskeho priestoru o 5-10 % v porovnaní s už existujúcimi riešeniami v oblasti automatizovanej detekcie strojovo-generovaných textov je tiež preukázateľne dosiahnuté v rámci štúdie CEaid [1] (V3.1, Obr. 1) porovnaním hodnôt metriky AUC ROC medzi najlepším existujúcim (state-of-the-art) detektorom Fast-DetectGPT a naším najlepším dotrénovaným detektorom Llama-3.2-3B (de-hu-pl). Fast-DetectGPT dosiahol v priemere hodnotu AUC ROC 0,7830, pričom náš detektor Llama-3.2-3B (de-hu-pl) dosiahol 0,9749, čo predstavuje relatívne zvýšenie o 24,5 %.

5 Kľúčové ukazovatele komunikácie, diseminácie a exploitácie

Tieto ukazovatele sú podrobne zdokumentované v samostatnej správe V2.1 Správa o výsledkoch komunikácie, diseminácie a exploitácie. V tejto časti uvádzame len sumarizáciu KPI vedeckých článkov ako primárneho diseminačného výstupu projektu.

Počtom vedeckých článkov sme prekročili plánovaný počet (plánované ≥ 3 , dosiahnuté ≥ 4), a teda plán je **splnený**. Počet vedeckých článkov uvedený vyššie zahŕňa minimálnu hodnotu tohto ukazovateľa dosiahnutú aj pri rovnomernom rozpočítaní príspevku projektu RobIndAI k danej publikácii pri článkoch obsahujúcich poďakovanie viacerým projektom. Teda ak článok obsahuje poďakovanie 2 projektom, do tejto hodnoty je započítaný len podiel 0,5. Celkový počet vedeckých článkov súvisiacich s projektom RobIndAI a obsahujúcich príslušné predpísané poďakovanie je teda vyšší (zoznam v Tab. 5).

Fórum	Nadpis	Stav	Link
CLEF 2025	mdok of KInIT: Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection	publikované	link
magazín Computer vydávaný IEEE Computer Society	Beyond Speculation: Measuring the Growing Presence of Large Language Model-Generated Texts in Multilingual Disinformation	publikované	DOI
kapitola knihy vydanej Springer Nature	False Alarm or Real Threat? Trends in GenAI-Mediated Disinformation	publikované	DOI
EMNLP 2025	A Rigorous Evaluation of LLM Data Generation Strategies for Low-Resource Languages	publikované	DOI
ACL 2026	Authorship Attribution in Multilingual Machine-Generated Texts	publikované	link
Findings of ACL 2026	CEAID: Benchmark of Multilingual Machine-Generated Text Detection Methods for Central European Languages	publikované	link
SemEval 2026	mcdok at SemEval-2026 Task 13: Finetuning LLMs for Detection of Machine-Generated Code	akceptované	preprint
EMNLP 2026	Increasing the Robustness of the Fine-tuned Multilingual Machine-Generated Text Detectors	v recenznom konaní	preprint
EMNLP 2026	Interpretable Predictability-Based AI Text Detection: A Replication Study	v recenznom konaní	preprint

Tab. 5 Zoznam zverejnených vedeckých článkov.

4 Referencie

- [1] Dominik Macko and Jakub Kopál. 2026. [CEAID: Benchmark of Multilingual Machine-Generated Text Detection Methods for Central European Languages](#). In Findings of the Association for Computational Linguistics: ACL 2026, pages 29177–29190, San Diego, California, United States. Association for Computational Linguistics.
- [2] Dominik Macko, Robert Moro, Ivan Srba. 2025. [Increasing the Robustness of the Fine-tuned Multilingual Machine-Generated Text Detectors](#). arXiv preprint arXiv:2503.15128.