

Monitorovacia správa - záverečná

Poradové číslo správy:	0002
Názov prijímateľa:	Kempelenov inštitút inteligentných technológií
Kód výzvy:	09I01-03-V04
Komponent	K09 Efektívnejšie riadenie a posilnenie financovania výskumu, vývoja a inovácií
Investícia	1. Podpora medzinárodnej spolupráce a zapájania sa do projektov Horizont Európa a EIT
Kód projektu:	09I01-03-V04-00059
Názov projektu	Robustnosť indikátorov dezinformačného obsahu vo viacjazyčnom online priestore
Meno a priezvisko hlavného riešiteľa:	prof. Ing. Mária Bieliková, PhD.
Monitorovacie obdobie (od do):	1.9.2025 - 30.6.2026
Typ monitorovacej správy:	záverečná

Popis činností uskutočnených v rámci projektu počas monitorovacieho obdobia.

Projekt aj vo svojej druhej (a finálnej) etape (v trvaní 10 mesiacov, v období od 1.9.2025 do 30.6.2026) postupoval v súlade s harmonogramom projektu. Dosiahnuté výsledky naplnili všetky stanovené ciele projektu (viď časť 6 tejto správy) a zabezpečili úspešné ukončenie projektu.

Druhá etapa projektu bola taktiež zameraná na široké spektrum predovšetkým výskumných, ale aj diseminačných a iných podporných aktivít.

Výskumné aktivity v pracovnom balíku NV3 zahŕňali predovšetkým detekciu strojovo-generovaného textu, ktoré priamo nadviazali na predchádzajúcu fázu. Porovnávacia štúdia viacjazyčnej detekcie strojovo-generovaných textov v stredoeurópskych jazykoch, ktorá bola reportovaná v priebežnej monitorovacej správe, bola dokončená a zodpovedajúci výskumný článok v rámci diseminačných aktivít pracovného balíka NV2 úspešne publikovaný na prestížnej konferencii ACL 2026 (sekcia Findings, top NLP konferencia s hodnotením CORE A*). Zodpovedajúci dataset CEALD, ako aj dataset doplnený o útočné vzorky boli publikované v anonymizovanej forme na platforme Zenodo.

Ďalšie výskumné aktivity zahŕňali zvyšovanie robustnosti detektorov strojovo-generovaného textu voči tzv. out-of-distribution dátam (t.j. textom, ktoré narušali distribúciu textov tréningovej sady), v rámci čoho sme vyvinuli metodiku dotrénovania robustných viacjazyčných detektorov. Tieto detektory ponúkajú vyššiu odolnosť voči neočakávaným dátam, ako napr. novým textovým doménam, príp. útokom na zabránenie detekcie. Metodiku sme tiež overili na zdieľanej úlohe workshopu SemEval 2026, ktorá bola zameraná na detekciu zdrojového kódu generovaného pomocou veľkých jazykových modelov. V podúlohe zameranej na binárnu detekciu dosiahol náš systém 88. percentil, čím potvrdil prenositeľnosť navrhnutej metodiky aj do iných domén (generovaný text vs. zdrojový kód). Výskumný článok v rámci tejto zdieľanej úlohy bol prijatý na workshop SemEval 2026. Výskumný článok zahŕňajúci metodiku robustného tréningu viacjazyčných detektorov strojovo-generovaného textu je v štádiu posudzovania na konferenciu EMNLP 2026 (CORE A*). Najlepšie modely detektorov boli publikované na platforme HuggingFace s prístupom obmedzeným na výskumné účely a zodpovedajúce zdrojové kódy na platforme GitHub s otvoreným prístupom.

V poslednej fáze zahŕňali výskumné aktivity skúmanie možností optimalizácie architektúry detekčného systému v stredoeurópskych jazykoch. V rámci porovnávacej štúdie sme porovnali tri architektúry detekčného systému: a) jedného viacjazyčne dotrénovaného modelu, b) kombinácie jednojazyčne dotrénovaných modelov pomocou priemerovania pravdepodobností predikcie alebo väčšinovej voľby a c) kombinácie jednojazyčne dotrénovaných modelov pomocou preferencie jazykovo korešpondujúceho detektora. Posledná z uvedených architektúr predstavuje najvhodnejšiu pre nasadenie detekčného systému v stredoeurópskom multilingválnom prostredí. Táto fáza ukončila výskumné aktivity plne napíňajúce všetky tri stanovené ciele projektu. V rámci NV3 boli tiež pripravené výstupy vo forme výskumných správ V3.1 a V3.2, čím boli tiež plne naplnené míľniky ML2 a ML3 (ML1 už bol naplnený v priebežnej monitorovacej správe).

Ako už bolo spomenuté vyššie, v rámci pracovného balíka NV2 pravidelne prebiehali diseminačné, komunikačné a exploitačné aktivity (napr. účasť na konferenciách, publikácia výskumných článkov, komunikácia na sociálnych sieťach, zverejnenie datasetov, modelov a zdrojových kódov), ktoré boli zosumarizované do záverečného reportu (V2.1). Tieto aktivity prebiehali zväčša podľa plánu - mierne oneskorenie publikácie výskumných výsledkov ku koncu projektu a publikovanie bez prezentácie (napr. v časopise) spôsobili mierny deficit v plánovaných odborných prezentáciách výsledkov projektu, ktorý sme ale úspešne dobehli prezentovaním výsledkov na lokálnych fórach ako HPC User Day a Slovak NLP.

V rámci podporných aktivít pracovného balíka NV1 počas monitorovaného obdobia prebiehalo riadenie projektu a boli kontinuálne vylepšované pôvodne nastavené procesy riadenia (zabezpečenia kvality, manažmentu rizík, manažmentu úloh, atď.). Bola tiež zabezpečovaná synergia so sesterským Horizon Europe projektom VIGILANT až do jeho konca v M13 riešenia projektu RobIndAI. V rámci tohto balíka boli pripravené ako priebežná (V1.1), tak aj táto záverečná monitorovacia správa (V1.2), čo predstavuje úplné naplnenie aj posledného míľnika projektu ML4.

Míľniky

Poradové číslo a názov míľnika	Typ a číslo monitorovacej správy	Názov pracovného balíka/pracovných balíkov	Popis dosiahnutia míľnika
ML1. Prvé verzie metód a modelov	priebežná, 0001	NV1 - Manažment	Míľnik dosiahnutý úplne. Nastavené procesy riadenia projektu. Výskum metód detekcie strojovo-generovaných textov viedol k prvým verziám metód a

			modelov (spolu so zodpovedajúcimi dátovými vzorkami a ďalšími podpornými metódami a modelmi). Názov dokumentov: Priebežná monitorovacia správa (V1.1)
ML2. Hotové metódy detekcie strojovo-generovaných textov	záverečná, 0002	NV3 - Výskum	Míľnik dosiahnutý úplne. Výskum metód detekcie strojovo-generovaných textov viedol k vyhodnoteným verziám metód a modelov aj vzhľadom na robustnosť voči útokom (spolu so zodpovedajúcimi dátovými vzorkami a ďalšími podpornými metódami a modelmi). Názov dokumentov: Výskumná správa o modeloch a metódach robustnej detekcie strojovo-generovaného textu (V3.1)
ML3. Hotová optimalizovaná architektúra detekčného systému	záverečná, 0002	NV3 - Výskum	Míľnik dosiahnutý úplne. Výskum optimalizácie architektúry viacjazyčného detekčného systému pre strojovo-generované texty viedol k identifikácii optimálneho riešenia vzhľadom na vyváženie rôznych požiadaviek (výkonnosť vs. výpočtová náročnosť). Názov dokumentov: Výskumná správa o optimalizovanej architektúre systému na detekciu strojovo-generovaného textu (V3.2)
ML4. Ukončenie projektu	záverečná, 0002	NV1 - Manažment NV2 - Impakt	Míľnik dosiahnutý úplne. Finalizácia riadiacich, diseminačných, komunikačných a exploitačných aktivít a príprava zodpovedajúcich dokumentov a reportov. Názov dokumentov: Záverečná monitorovacia správa (V1.2) Správa o výsledkoch komunikácie, diseminácie a exploitácie (V2.1)

Výstupy projektu

Poradové číslo a názov výstupu	Typ a číslo monitorovacej správy	Názov pracovného balíka/pracovných balíkov	Popis dosiahnutia výstupu
1. V1.1 Priebežná správa o implementácii a dosiahnutých výsledkoch projektu	priebežná, 0001	NV1 - Manažment	Výstup dosiahnutý úplne. Výstupom je priebežná monitorovacia správa dokumentujúca stav aktivít na konci M10. Názov dokumentov: Priebežná monitorovacia správa (V1.1)
2. V1.2 Záverečná správa o dosiahnutých výsledkoch projektu	záverečná, 0002	NV1 - Manažment	Výstup dosiahnutý úplne. Výstupom je táto záverečná monitorovacia správa dokumentujúca stav aktivít na konci M20. Názov dokumentov: Záverečná monitorovacia správa (V1.2)
3. V2.1 Správa o výsledkoch komunikácie, diseminácie a exploitácie	záverečná, 0002	NV2 - Impakt	Výstup dosiahnutý úplne. Výstupom je správa dokumentujúca výsledky aktivít komunikácie, diseminácie a exploitácie na konci M20. Názov dokumentov: Správa o výsledkoch komunikácie, diseminácie a exploitácie (V2.1)

4. V3.1 Výskumná správa o modeloch a metódach robustnej detekcie strojovo-generovaného textu	záverečná, 0002	NV3 - Výskum	Výstup dosiahnutý úplne. Výstupom je správa dokumentujúca výsledky výskumu modelov a metód robustnej detekcie strojovo-generovaného textu na konci M16. Názov dokumentov: Výskumná správa o modeloch a metódach robustnej detekcie strojovo-generovaného textu (V3.1)
5. V3.2 Výskumná správa o optimalizovanej architektúre systému na detekciu strojovo-generovaného textu	záverečná, 0002	NV3 - Výskum	Výstup dosiahnutý úplne. Výstupom je správa dokumentujúca výsledky výskumu optimalizácie architektúry systému na detekciu strojovo-generovaného textu na konci M20. Názov dokumentov: Výskumná správa o optimalizovanej architektúre systému na detekciu strojovo-generovaného textu (V3.2)

Identifikácia problémov a rizík

Zdôvodnenie odchýlok:	V rámci žiadosti boli identifikované relevantné riziká na celú dobu riešenia projektu a boli navrhnuté opatrenia na ich minimalizáciu. Môžeme konštatovať, že v rámci riešenia projektu sa žiadne z rizík nenaplnilo a riešenie projektu prebieha podľa pôvodného plánu. Zároveň konštatujeme, že neboli identifikované ani žiadne nové riziká, s ktorými by pôvodná projektová žiadosť nepočítala.
-----------------------	--

Zhodnotenie naplnenia cieľ'a projektu

--

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:

Cieľ 1: Detekcia strojovo generovaného textu

Cieľ 2: Zvýšenie robustnosti detekčných metód

Cieľ 3: Výber vhodnej architektúry

Cieľ 1 bol naplnený prostredníctvom realizácie porovnávacej štúdie CEAID 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. Štúdia priniesla 5 špecializovaných AI detektorov dotrénovaných na vzorkách textu dedikovaných pre konkrétny jazyk, ako aj množstvo detektorov dotrénovaných na kombináciách týchto piatich tréningových jazykov. Vzhľadom na veľké množstvo detektorov, tieto boli zverejnené prostredníctvom zverejnenia tréningových dát (CEAID dataset na platforme Zenodo) a tréningových zdrojových kódov umožňujúcich plnú replikáciu modelov (GitHub).

Cieľ 2 bol naplnený prostredníctvom návrhu škálovateľnej metodiky tréningovania robustných viacjazyčných detektorov založených na jazykových modeloch, ako aj prostredníctvom rozšírenia datasetu CEAID o útočné vzorky pomocou homoglyfného útoku a parafrázovania. 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. Oba datasety, pôvodný CEAID aj rozšírenie o útočné vzorky boli publikované v anonymizovanej forme na platforme Zenodo. Najlepší model vyplývajúci zo štúdie predstavenia metodiky tréningovania robustných viacjazyčných detektorov je publikovaný na platforme HuggingFace.

Cieľ 3 bol naplnený prostredníctvom realizácie porovnávacej štúdie zameranej na porovnanie troch rôznych možností realizácie viacjazyčného detekčného systému, 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.

Projekt RobIndAI prispel k riešeniu jednej z kľúčových výziev súčasného online priestoru, a to schopnosti spoľahlivo identifikovať strojovo-generovaný text ako indikátor potenciálneho dezinformačného obsahu. Oproti existujúcemu stavu výskumu (napr. v projekte VIGILANT) projekt priniesol pridanú hodnotu najmä v pokrytí stredoeurópskych jazykov a dvoch textových domén (novinové články a sociálne médiá), ktorým sa medzinárodný výskum venuje len okrajovo, v systematickom zohľadnení odolnosti detekčných metód voči útokom (homoglyfy, parafrázovanie), ako aj navrhnutou metodikou tréningovania robustných viacjazyčných detektorov. Okrem toho projekt tiež prispel k poznatkom o vplyve kombinácií tréningových jazykov na medzijazykovú prenositeľnosť detekčných schopností a identifikáciou optimálnej architektúry viacjazyčného detekčného systému. Otvorené zverejnenie datasetov, modelov, ako aj zdrojových kódov umožňuje ich ďalšie využitie výskumnou komunitou aj aplikačnou praxou.

Výsledky projektu sú dostupné odbornej verejnosti prostredníctvom publikácií (väčšina dostupná vo forme otvoreného prístupu), datasetov (Zenodo), modelov (HuggingFace) a zdrojových kódov (GitHub). Vytvorené modely, metódy a metodiky sú pripravené na nasadenie v podporných systémoch mediálnych profesionálov a policajných autorít. Priebežné výsledky projektu sa tiež pretavili do podania nového návrhu európskeho grantového projektu v schéme Marie Curie Sklodowska Actions, ako aj do podania nového projektu Horizon Europe FIMI-RESIST. V rámci sesterského Horizon Europe projektu VIGILANT bola v prototypu MVP použitá verzia modelu dotrénovaná pomocou robustnej metodiky tréningovania viacjazyčných detektorov, vyvinutá v projekte RobIndAI. V krátkodobom horizonte (bezprostredne po skončení projektu) sa očakáva v Horizon Europe projekte AI-CODE využitie poznatkov a priame využitie metodiky tréningovania robustných detektorov na tréningovanie novej verzie detekčného modelu neviazaného na dáta použiteľné len na výskum (t. j. detektor potenciálne využiteľný aj v komerčných nástrojoch pre profesionálov).

S ohľadom na priame využitie výsledkov RobIndAI v prototypu projektu VIGILANT, ale aj berúc do úvahy diskusie so zástupcami DG CONNECT Európskej komisie, komunikáciu odbornej verejnosti prostredníctvom webináru Horizon Europe projektu vera.ai, ako aj ostatné účasti RobIndAI na odborných fórach a okrúhlych stoloch, sa v strednodobom horizonte (1-2 roky po skončení projektu) predpokladá postupná integrácia metód a modelov do bežných pracovných postupov vyšetrovateľov policajných autorít, ako aj mediálnych profesionálov (napr. na overovanie faktov), a ich ďalší rozvoj zo strany KInIT aj iných výskumných pracovísk. V dlhodobom horizonte (viac ako 3 roky po skončení projektu) môžu detekčné mechanizmy vyvinuté v projekte RobIndAI prispieť k tomu, že spoľahlivá detekcia strojovo-generovaného textu sa stane štandardnou súčasťou online médií, čím sa skomplikuje zneužívanie generatívnych jazykových modelov na šírenie dezinformácií a posilní dôvera verejnosti v informačné zdroje.

Výstupy projektu budú ďalej rozvíjané v nadväzujúcich výskumných aktivitách KInIT (napr. spomínané priame nadviazanie v projekte AI-CODE), vrátane rozširovania jazykového pokrytia a adaptácie metód na nové typy útokov a nové generatívne modely. Datasetsy, modely, ako aj zdrojové kódy sú zverejnené s otvoreným prístupom a môžu byť znovu použité pre príbuzné aplikácie, ako je detekcia plagiatov, či odhaľovanie automatizovaných podvodov. Výsledky projektu budú naďalej prezentované na medzinárodných konferenciách (napr. na ACL 2026 konajúcej sa po oficiálnom ukončení projektu RobIndAI) a v odborných publikáciách.

Fotodokumentácia k preukázaniu zabezpečenia publicity projektu: dostupné v prílohe Správa o výsledkoch komunikácie, diseminácie a exploitácie (V2.1).
Dokumentácia preukazujúca dosiahnutú hodnotu merateľných výsledkoch projektu/naplnenia míľnikov a cieľa projektu po realizácii projektu: dostupné v prílohe Správa o dosiahnutých merateľných ukazovateľoch (Príloha A).
Dokumentácia preukazujúca splnenie príslušných zmluvných podmienok určených v konkrétnej ZoPPM a vzťahujúcich sa k ukončeniu realizácie aktivít projektu: dostupné v tomto dokumente Záverečná monitorovacia správa (V1.2).
Dokumentácia preukazujúca splnenie cieľov/míľnikov projektu a realizovaných opatrení v rámci aktivít projektu (v zmysle ZoPPM a príslušnej výzvy): dostupné v prílohe Správa o dosiahnutých merateľných ukazovateľoch (Príloha A).

Zoznam príloh

Prílohy

	názov <u>V1.1 Priebežná správa o implementácii a dosiahnutých výsledkoch projektu.pdf</u> popis Príloha 1: Priebežná monitorovacia správa (V1.1)
	názov <u>V2.1 Správa o výsledkoch komunikácie, diseminácie a exploitácie.pdf</u> popis Príloha 2: Správa o výsledkoch komunikácie, diseminácie a exploitácie (V2.1)
	názov <u>V3.1 Výskumná správa o modeloch a metódach robustnej detekcie strojovo-generovaného textu.pdf</u> popis Príloha 3: Výskumná správa o modeloch a metódach robustnej detekcie strojovo-generovaného textu (V3.1)
	názov <u>V3.2 Výskumná správa o optimalizovanej architektúre systému na detekciu strojovo-generovaného textu.pdf</u>



názov Príloha A Správa o dosiahnutých merateľných ukazovateľoch.pdf
popis Príloha 5: Správa o dosiahnutých merateľných ukazovateľoch (Príloha A)

Vyhlásenie za osobu zodpovednú za projekt/časť projektu

Ja, dolu podpísaná osoba zodpovedná za projekt/časť projektu čestne vyhlasujem, že údaje uvedené v tejto monitorovacej správe a všetkých jej prílohách sú úplné a pravdivé.

Meno a priezvisko osoby zodpovednej za projekt/časť projektu: prof. Ing. Mária Bieliková, PhD.

Dátum podpisu a podpis osoby zodpovednej za projekt/časť projektu: 23.7.2026

Vyhlásenie za prijímateľa

Ja, dolu podpísaný/á štatutárny orgán alebo osoba oprávnená konať za prijímateľa čestne vyhlasujem, že údaje uvedené v tejto monitorovacej správe a všetkých jej prílohách sú úplné a pravdivé.

Meno a priezvisko štatutárneho zástupcu/ osoby poverenej konať za prijímateľa: prof. Ing. Mária Bieliková, PhD.

Dátum podpisu štatutárneho zástupcu alebo osoby poverenej konať za prijímateľa: