Monitorovacia správa - priebežná

Poradové číslo správy:	**0001**
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):	**1.11.2024 - 31.8.2025**
Typ monitorovacej správy:	**priebežná**

Popis činností uskutočnených v rámci projektu počas monitorovacieho obdobia.

Popis:

Projekt po svojej prvej etape (v trvaní 10 mesiacov, v období od 1.11.2024 do 31.8.2025) postupuje v súlade s harmonogramom projektu. Doterajšie výsledky poskytujú dobrý predpoklad pre úspešné ukončenie projektu po jeho druhej polovici.

Prvá etapa projektu bola zameraná na široké spektrum predovšetkým výskumných, ako aj podporných aktivít.

Výskum zahŕňal rešerše dostupnej literatúry, analýzu a replikáciu kľúčových prác (získanie datasetov, sfunkčnenie dostupných metód generovania textu a detekcie stojovo-generovaného textu), zostavovanie benchmarkového datasetu zameraného na stredoeurópsky región, dizajn nových metód a modelov (najmä ich dotrénovanie pre účely detekcie) a ich experimentálne overovanie. Analyzovali sme tiež potreby cieľových skupín vytváraných metód a modelov. Výskumná aktivita sa vykonávala v rámci pracovného balíka NV3 a prispievala do všetkých troch výskumných úloh v ňom.

Hlavná časť výskumného úsilia sa sústredila na detekciu strojovo-generovaného textu, ktorý je potrebné odlíšiť od textu písaného človekom. Analýza existujúcich dátových sád identifikovala potrebu vytvorenia nového datasetu špecializovaného pre stredoeurópsky región (t.j. zahŕňajúci dostatočný počet vzoriek textov vo viacerých jazykoch tohto regiónu), ktorý by kvôli robustnosti evaluácie konzistentne pokrýval viaceré domény (napr. novinové články, sociálne siete) a generátory (veľké jazykové modely – LLM) spolu s dostatočným počtom vzoriek textov písaných človekom v daných jazykoch. Výskumné úsilie vyústilo do vytvorenia nového datasetu pokrývajúceho sedem jazykov stredoeurópskeho regiónu, pomocou kombinácie dostupných dátových vzoriek z existujúcich dátových sád.

Tento nový dataset bol použitý na realizáciu porovnávacej štúdie metód detekcie strojovo-generovaného textu v týchto jazykoch. Pristúpili sme k zmapovaniu a porovnaniu existujúcich metód detekcie troch kategórií – štatististické (využívajúce rozdieli medzi sledovanými triedami textov v štatistickej distribúcii rôznych ukazovateľov na základe reprezentácie poskytnutej pomocou LLM), predtrénované (jazykové modely trenované na úlohu detekcie na iných datasetoch) a dotrénované ("fine-tuned" modely, špeciálne dotrénované na trénovacej časti vytvoreného datasetu). Špecializované dotrénované modely poskytujú najvyššiu úspešnosť detekcie v stredoeurópskych jazykoch, konzistentne pre všetky generátory a domény. Súčasťou štúdie bolo porovnanie výberu jazykov (rôznych kombinácií) zahrnutých do dotrénovania detekčných metód a určenie ich vplyvu na výslednú úspešnosť detekcie. Metódy obsahujúce aspoň tri jazyky pri dotrénovaní boli vo všeobecnosti úspešnejšie ako metódy obsahujúci menší počet jazykov. Jedným z prínosov uvedenej štúdie sú aj spomínané jazykovo-špecializované modely pre každý z trénovacích jazykov, pokrývajúce aj jazyky, pre ktoré doteraz neexistovali špecializované detekčné modely (dané jazyky neboli zahrnuté do existujúceho výskumu). Táto štúdia spolu s vytvoreným datasetom priamo napĺňa cieľ 1 vytýčený v projektovom zámere.

Keďže projekt si kladie za cieľ zvyšovanie robustnosti metód detekcie stojovo-generovaných textov, vyššie opísanú dátovu sadu sme obohatili (data augmentation) pomocou vybraných techník zabránenia detekcie (parafrázovanie a homoglyfný útok). Obe tieto techniky bolo možné aplikovať multilingválne, konzistentne vo všetkých zvolených jazykoch stredoeurópskeho priestoru. Takto obohatená sada bola použitá na vyhodnotenie základnej odolnosti detekčných metód, ktorá bude slúžiť na referenčné porovnanie v ďalšej fáze projektu zameranej na zvýšenie odolnosti voči takýmto útokom. Robustnosť vyvinutých metód dotrénovania modelov bola navyše otestovaná v účasti na zdielanej úlohe Voight-Kampff Generative AI Detection v rámci PAN labu konferencie CLEF 2025. Navrhnuté riešenie bolo vyhodnotené ako najrobustnejšie voči útočným metódam zabránenia detekcie (1. miesto). Táto úloha bola síce zameraná len na anglické texty, ale vyvinuté riešenie je rozšíriteľné aj na iné jazyky. Obdobne robustne dotrénovaný multilingválny detektor sme použili na odhad prevalencie strojovo-generovaných textov vo vjacjazyčnom datasete MultiClaim, obsahujúcom príspevky zo sociálnych sietí overené profesionálnymi fact-checkermi (zväčša dezinformačný obsah). Táto štúdia prevalencie bola prijatá na publikáciu v prestížnom magazíne Computer, publikovanom IEEE Computer Society.

Výsledky doteraz uskutočneného experimentálneho výskumu, najmä detekčné modely a metódy (a spolu s nimi vytvorené dátové vzorky) v plnej miere napĺňajú prvý míľnik projektu (ML1 Prvé verzie metód a modelov M13).

Popri výskumnej činnosti prebiehala aj podporná činnosť. Niektoré z výstupov projektu boli počas prvej etapy publikované a prezentované na odborných a popularizačných fórach – účasť na zdieľanej úlohe a prezentácia najlepšieho riešenia na konferencii CLEF 2025 zameranej na otvorené viacjazyčné porovnávacie štúdie, ako aj akceptovaná publikácia v magazíne Computer. Priebežné výsledky projektu sa tiež pretavili do podania nového návrhu európskeho grantového projektu v schéme Marie Curie Sklodowska Actions. Komunikačné, diseminačné a exploitačné aktivity prebiehali v pracovnom balíku NV2. Na začiatku realizácie bola vytvorená webová stránka projektu (https:///kinit.sk/project/robindai-robust-indication-of-ai-generated-disinformation-content/), príprava príspevkov na sociálnych sieťach a newslettrov. V rámci podporných aktivít boli počas monitorovaného obdobia prebiehalo riadenie projektu a boli nastavené a kontinuálne vylepšované procesy tohto riadenia (zabezpečenia kvality, manažmentu rizík, manažmentu úloh, atď.). Bola tiež zabezpečovaná synergia so sesterským Horizon Europe projektom VIGILANT.

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:** **Táto monitorovacia správa (V1.1)** **1. Výskumný článok prijatý na konferenciu PAN@CLEF 2025** **2. Výskumný článok prijatý do magazínu Computer** **3. Výskumný článok prijatý na konferenciu EMNLP 2025** **4. Výskumný článok poslaný na konferenciu AAAI 2026**
ML2. Hotové metódy detekcie strojovo-generovaných textov	**x**	**NV3 - Výskum**	**Miľník nebol v danom monitorovacom období dosiahnutý.**
ML3. Hotová optimalizovaná architektúra detekčného systému	**x**	**NV3 - Výskum**	**Miľník nebol v danom monitorovacom období dosiahnutý.**
ML4. Ukončenie projektu	**x**	**NV1 - Manažment** **NV2 - Impakt**	**Miľník nebol v danom monitorovacom období dosiahnutý.**

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 táto monitorovacia správa dokumentujúca stav aktivít na konci M10.**

2. V1.2 Záverečná správa o dosiahnutých výsledkoch projektu	**x**	**NV1 - Manažment**	**Výstup nebol v danom monitorovacom období dosiahnutý.**
3. V2.1 Správa o výsledkoch komunikácie, diseminácie a exploitácie	**x**	**NV2 - Impakt**	**Výstup nebol v danom monitorovacom období dosiahnutý.**
4. V3.1 Výskumná správa o modeloch a metódach robustnej detekcie strojovogenerovaného textu	**x**	**NV3 - Výskum**	**Výstup nebol v danom monitorovacom období dosiahnutý.**
5. V3.2 Výskumná správa o optimalizovanej architektúre systému na detekciu strojovo-generovaného textu	**x**	**NV3 - Výskum**	**Výstup nebol v danom monitorovacom období dosiahnutý.**

Identifikácia problémov a rizík

Zdôvodnenie odchýlok:

Počas monitorovaného obdobia boli identifikované nasledujúce problémy/riziká aj s príslušným opisom ich riešenia:

Nedostatok konzistentných trénovacích a testovacích dát
Detekcia strojovo-generovaných textov je náchylná na nevyvážené množiny dát z hľadiska ich jazykového, doménového a tematického zastúpenia. Jazyky stredoeurópskeho regiónu boli slabo zastúpené v existujúcich dostupných datasetoch. Problém sme vyriešili kombináciou vzoriek z viacerých datasetov zahŕňajúciich viaceré domény pri zohľadnení totožných skupín generátorov textov. Vzhľadom na počet vzoriek sme vyčlenili dostatočný počet textov na evaluáciu (testovacia časť datasetu) pre všetky jazyky stredoeurópskeho regiónu a vybrali len podskupinu jazykov s dostatočným počtom vzoriek na trénovanie (vyvážených skrz generátory, jazyky aj domény).

Robustnosť, zovšeobecniteľnosť a škálovateľnosť AI metód
Skúmané metódy detekcie strojovo-generovaných textov môžu vykazovať problémy s ich robustnosťou, zovšeobecniteľnosťou a škálovaním. Ako súčasťou mitigácie v tomto smere sme sa sústredili na efektívne trénovanie AI modelov s obmedzeným množstvom dát a prostriedkov (angl. learning with limited labelled data), skúmanie citlivosti modelov na náhodnosť v dátach, a tiež na možnosti augumentácie dát (špeciálne o útočné vzorky obmedzujúce detekciu). Robustnosť použitej metódy dotrénovania AI modelov sme overili na nezávislej zdielanej úlohe zameranej na robustnosť detekčných modelov (použitá metóda obstála najlepšie).

Etické a legálne aspekty projektu
Vzhľadom na viaceré aspekty výskumu realizovanom v projekte (predovšetkým súvis s DSA, použitie AI metód, implementácia automatizovaných agentov, analýza šírenia problematického a škodlivého obsahu), neoddeliteľnou súčasťou realizácie projektu je aj vysporiadanie sa s príslušnými etickými a legálnymi aspektmi (vrátane EU legislatívy, ako je DSA, AIA, GDPR). Mitigačná stratégia zahŕňa priame zapojenie expertov na etické a legálne aspekty ako súčasť tímu pracujúcom na projekte a ich úzku koordináciu s výskumným tímom.

Popis plánovaných činností v nasledujúcom monitorovacom období

Popis plánovaných činností:	**V nasledujúcom monitorovacom období (v trvaní 10 mesiacov, od 1.9.2025 do 30.6.2026) budú prebiehať predovšetkým nasledujúce činnosti:** **1. Pokračujúci výskum robustných metód detekcie strojovo-generovaného textu** **Nadviažeme na existujúce prototypy metód a modelov a budeme pracovať na ich zdokonaľovaní vzhľadom na presnosť, výpočtovú efektívnosť, a najmä robustnosť voči útokom. Špecificky sa viac zameriame na robustný spsob dotrénovania modelov s využitím obohatených trénovacích vzoriek (obsahujúcich samotné útočné vzorky). Zvýšenie robustnosti voči útokom overíme na už vytvorenej obohatenej vzorke viacjazyčných textov a porovnáme s už vyhodnotenou základnou mierou robustnosti detekčných metód obsiahnutou v realizovanej porovnávacej štúdii.** **2. Výskum optimalizácie architekúry detekčného systému** **Vytvorené dataset stredoeurópskych jazykov použijeme na skúmanie a identifikáciu optimalizovanej architektúry detekčného systému. Porovnéme kombináciu (ensemble) jednojazyčne dotrénovaných modelov s najlepším viacjazyčne dotrénovaným modelom z predchádzajúcej porovnávacej štúdie. Evaluácia bude prebiehať na už vytvorenej testovacej dátovej sade.** **3. Manažment projektu, diseminácia, komunikácia, exploitácia a klustering** **Bude sa pokračovať v manažmente projektu na exekutívnej dennej ako aj strednodobej strategickej báze. S pribúdajúcimi výsledkami sa rozšíria aktivity zameraná na potenciálnu exploitáciu riešenia a zabezpečenie jeho udržateľnosti. Členovia tímu sa zúčastnia medzinárodných aj národných fór, na ktorých budú osobne prezentovať výsledky projektu (napr. realizovanú porovnávaciu štúdiu pre stredoeurópske jazyky). Budeme naďalej pokračovať v propagácií projektu a dosiahnutých výsledkov, a to ako pre širokú verejnosť (napr. na Európskej noci vedy 2025), tak aj v rámci vedeckej komunity a stakeholderov.**

Zoznam príloh

Prílohy

Title	**4. Článok poslaný na AAAI 2026 Authorship Attribution in Multilingual Machine-Generated Texts.pdf**
Description	

Title	**3. Článok prijatý na EMNLP 2025 A Rigorous Evaluation of LLM Data Generation Strategies for Low-Resource Languages.pdf**
Description	

Title	**2. Článok prijatý do magazínu Computer Beyond speculation Measuring the growing presence of LLM-generated texts in multilingual disinformation.pdf**
Description	

Title	**1. Článok prijatý na PAN@CLEF 2025 Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection.pdf**
Description	

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:	**29.9.2025**

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:	**29.9.2025**

Name and surname of person completing the form	**Katarína Házyová**
	✓ I confirm that information provided in this application is true and correct.