V3.1 Výskumná správa o modeloch a metódach robustnej detekcie strojovo-generovaného textu

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

Úvod

Vzhľadom na schopnosť moderných jazykových modelov generovať vysokokvalitný text v rôznych jazykoch, ktorý je pre človeka nerozoznateľný, je obava zo zneužitia tejto technológie rastúca (napr. medzinárodné dezinformačné kampane). Spoľahlivá detekcia strojovo-generovaného textu a jeho rozlíšenie od originálneho textu písaného človekom je v tomto ohľade podstatným a veľmi žiadaným indikátorom.

Strojovo-generovaný text (MGT – angl. machine-generated text) v rámci nášho výskumu predstavuje text vygenerovaný alebo výrazne modifikovaný jazykovými modelmi umelej inteligencie (AI), zväčša tzv. veľkými jazykovými modelmi (LLM – angl. large language model). Taktiež zameranie na detekčné metódy je orientované výlučne na metódy založené na AI.

Táto výskumná správa je rozdelená na tri časti. V prvej časti (Kap. 2.1) sa zameriavame na porovnávaciu štúdiu MGT detekčných metód pre jazyky stredoeurópskeho regiónu. V druhej časti (Kap. 2.2) sa zameriavame na zvýšenie robustnosti metód detekcie MGT. Tretia časť (Kap. 2.3) sumarizuje aplikáciu robustných metód detekcie MGT pre odhad prevalencie MGT v dezinformačnom priestore ako aj modifikáciu na detekciu konkrétneho generátora MGT.

2 Modely a metódy robustnej detekcie MGT

V našom výskume sa zameriavame ako na porovnanie rôznych metód a prístupoch detekcie MGT založených na jazykových AI modeloch, tak aj na zvýšenie robustnosti detektorov pre ich praktické využitie (vo forme odolnosti voči útokom, či aplikácii na MGT vygenerované novými LLM).

2.1 Detekcia MGT v stredoeurópskych jazykoch

Existujúce metódy detekcie MGT boli natrénované na jazykoch s veľkým množstvom zdrojov (najmä angličtina, ale aj ruština, či španielčina), a preto ich aplikácia pre ostatné jazyky dosahuje nižšiu úspešnosť (spoliehajúc sa na transfer schopností kvôli obrovskému množstvu dát v rámci predtrénovania jazykových modelov). Špecializované detekčné metódy pre menšie jazyky chýbajú. V rámci tejto porovnávacej štúdie [1] sme za zamerali na jazyky stredoeurópskeho regiónu, ako je definovaný podľa Bideleux and Jeffries (2007), konkrétne češtinu, chorvátčinu, maďarčinu, nemčinu, poľštinu, slovenčinu a slovinčinu. Nielenže sme overili úspešnosť detekcie MGT v týchto jazykoch, ale porovnali sme aj texty z domény formálnych novinových článkov s doménou sociálnych médií, vyhodnotili sme vplyv kombinácie trénovacích jazykov z tejto množiny na výslednú detekciu a dotrénovali sme špecializované detekčné modely pre tieto jazyky.

Dataset pre experimenty v rámci tejto porovnávacej štúdie sme vytvorili kombináciou dát v týchto jazykoch z našich existujúcich datasetov MULTITuDE_v3 (novinové články) a MultiSocial (texty zo sociálnych médií). Tieto dáta obsahujú texty písané ľuďmi (získané zo starších zdrojov), ako aj texty k nim vygenerované pomocou 8 rôznych LLM: Aya-101, Gemini (len sociálne médiá), GPT-3.5-Turbo-0125, Llama-2-70B-chat-hf (len novinové články), Mistral-7B-Instruct-v0.2, OPT-IML-Max-30B, v5-Eagle-7B-HF, and Vicuna-13B. Texty v slovenskom a slovinskom jazyku sú použité len na testovanie, keďže neobsahovali dostatočný počet trénovacích vzoriek pre doménu sociálnych médií. Z hľadiska metód detekcie MGT sme do štúdie zahrnuli metódy z troch kategórií: štatistické metriky (Binoculars, Fast-DetectGPT, LLM-Deviation), predtrénované metódy (ChatGPT-detector-RoBERTa-Chinese, Detection-Longformer, BLOOMZ-3B-mixed-detector), a metódy dotrénované v tejto štúdii (mDeBERTa-v3-base, XLM-RoBERTa-base, Llama-3.2-3B, Gemma-2-2B). Na evaluáciu sme zvolili metriku AUC ROC, ktorá reprezentuje všeobecnú

detekčnú schopnosť (bez stanovenia konkrétneho prahu pre klasifikáciu medzi dvomi triedami). Jazyky sú označené dvojpísmenovým kódom podľa ISO 639-1.

Výsledky štúdie na Obr. 1 ukázali, že dotrénované špecializované MGT detektory dosahujú omnoho vyššiu úspešnosť detekcie ako existujúce detektory zvyšných dvoch kategórií vo všetkých testovacích jazykoch. Najnižšiu úspešnosť detekcie dosiahli predtrénované detekčné modely, pre ktoré je zvolená množina stredoeurópskych jazykov pravdepodobne príliš odlišná od ich trénovacích dát. O niečo vyššiu úspešnosť dosiahli štatistické detektory, ktorým túto schopnosť zabezpečilo použitie viacjazyčného mGTP LLM na pozadí. Doména novinových článkov je jednoduchšia na detekciu ako sociálne médiá, s výnimkou dvoch predtrénovaných detektorov, ktoré boli zjavne trénované na kratších textoch. Poľština a slovinčina sa celkovo javia ako najťažšie na detekciu.

Domain	Detector	All	cs	de	hr	hu	pl	sk	sl
News	Llama-3.2-3B (hr-hu-cs)	**0.9952**	0.9976	**0.9926**	**0.9994**	**0.9967**	**0.9937**	0.9928	0.9943
	mDeBERTa-v3-base (cs)	0.9940	**0.9986**	0.9921	0.9924	0.9925	0.9900	**0.9981**	**0.9973**
	Gemma-2-2B (de-pl-cs)	0.9911	0.9966	0.9912	0.9827	0.9882	0.9878	0.9978	0.9894
	XLM-RoBERTa-base (de-pl-hr-hu-cs)	0.9824	0.9896	0.9728	0.9876	0.9759	0.9847	0.9769	0.9910
	Fast-DetectGPT	0.8490	0.8773	0.8717	0.8777	0.7867	0.8351	0.8413	0.9090
	Binoculars	0.8341	0.8771	0.8536	0.8707	0.7809	0.8228	0.8298	0.8965
	LLM-Deviation	0.7060	0.9083	0.7298	0.9025	0.6429	0.7507	0.8048	0.9072
	Detection-Longformer	0.6503	0.6356	0.6074	0.7595	0.7507	0.7003	0.4962	0.7168
	ChatGPT-detector-RoBERTa-Chinese	0.6223	0.5364	0.7168	0.6672	0.6717	0.6541	0.7646	0.7038
	BLOOMZ-3B-mixed-detector	0.5626	0.5049	0.5963	0.5271	0.4680	0.5691	0.6970	0.5544
Social media	Llama-3.2-3B (de-pl-hu)	**0.9506**	**0.9800**	0.9427	**0.9628**	**0.9744**	**0.9466**	**0.9527**	0.9176
	mDeBERTa-v3-base (de)	0.9476	0.9536	**0.9515**	0.9503	0.9708	0.9413	0.9439	**0.9306**
	XLM-RoBERTa-base (de-pl)	0.9412	0.9590	0.9344	0.9548	0.9670	0.9282	0.9424	0.8972
	Gemma-2-2B (de-pl-hr-hu-cs)	0.9313	0.9631	0.9334	0.9468	0.9686	0.9240	0.9324	0.8483
	LLM-Deviation	0.8049	0.8877	0.7279	0.8030	0.8990	0.8128	0.7818	0.7484
	Binoculars	0.7699	0.7911	0.7922	0.8107	0.7856	0.7598	0.7384	0.7169
	BLOOMZ-3B-mixed-detector	0.7627	0.7989	0.7843	0.7748	0.8394	0.7661	0.7625	0.6438
	Fast-DetectGPT	0.7617	0.7626	0.7827	0.8044	0.7780	0.7500	0.7467	0.7238
	ChatGPT-detector-RoBERTa-Chinese	0.6737	0.6605	0.7974	0.6211	0.7778	0.6179	0.6345	0.6480
	Detection-Longformer	0.4757	0.5054	0.3848	0.5480	0.5288	0.4772	0.4293	0.4382

Obr. 1 Porovnanie detekčných schopností (AUC ROC) pre jednotlivé testovacie jazyky a domény. Pri dotrénovaných detektoroch je v zátvorke uvedená kombinácia trénovacích jazykov dosahujúca najlepší výsledok [1].

Zaujímavé je, že žiadna z kombinácií trénovacích jazykov pre najlepšie dotrénované metódy nie je rovnaká medzi dvomi doménami. Preto sme sa detailnejšie pozreli na všetky kombinácie trénovacích jazykov a zosumarizovali výsledky (priemer medzi 4 základnými modelmi) na Obr. 2. Výsledky indikujú, že nemčina a poľština sú dôležité pri trénovaní (oba jazyky obsiahnuté v 7 kombináciách z najlepších 10). Je tiež dôležité zahrnúť všeobecne aspoň dva jazyky do trénovania, keďže všetky jednojazyčné kombinácie sú v najhorších 10.

Taktiež 8 z najlepších 10 kombinácií obsahuje aspoň 3 jazyky. Generalizácia do slovinčiny sa zdá byť najťažšia, pričom si vyžaduje poľštinu alebo češtinu v trénovacej kombinácii, aby bola detekcia najlepšia. Chorvátčina je v transferabilite do slovinčiny najhoršia, hoci predstavuje jazyk susedného štátu a patrí do spoločnej vetvy jazykovej rodiny.

Train	**All**		**cs**		**de**		**hr**		**hu**		**pl**		**sk**		**sl**	
Languages	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
cs-de-hr-hu-pl	**0.967**	0.005	0.982	0.003	0.963	0.010	0.975	0.002	0.979	0.004	0.962	0.007	0.969	0.005	0.941	0.016
de-hu-pl	0.966	0.009	**0.982**	0.007	0.967	0.011	0.965	0.012	0.980	0.005	0.966	0.010	0.969	0.011	**0.953**	0.014
de-pl	0.966	0.008	0.981	0.006	0.969	0.009	0.965	0.012	0.974	0.011	0.967	0.006	0.967	0.011	0.951	0.010
cs-de	0.966	0.004	0.980	0.002	**0.972**	0.007	0.963	0.012	0.973	0.008	0.951	0.007	**0.973**	0.006	0.951	0.004
cs-de-pl	0.966	0.006	0.981	0.005	0.966	0.009	0.961	0.014	0.975	0.009	0.962	0.007	0.971	0.008	0.948	0.009
cs-de-hr-pl	0.965	0.008	0.981	0.004	0.965	0.010	0.974	0.005	0.973	0.007	0.962	0.008	0.966	0.010	0.945	0.017
cs-hr-pl	0.962	0.007	0.982	0.004	0.948	0.016	0.975	0.003	0.977	0.004	0.966	0.005	0.965	0.011	0.935	0.015
cs-de-hu-pl	0.962	0.008	0.980	0.008	0.964	0.011	0.957	0.021	0.973	0.005	0.959	0.009	0.966	0.013	0.945	0.014
de-hr-pl	0.962	0.004	0.980	0.005	0.963	0.010	0.973	0.008	0.974	0.006	0.959	0.009	0.972	0.004	0.932	0.012
cs-de-hr-hu	0.961	0.006	0.978	0.003	0.964	0.012	0.975	0.005	0.978	0.003	0.950	0.008	0.961	0.007	0.929	0.016
cs-hu-pl	0.961	0.009	0.981	0.006	0.946	0.013	0.968	0.009	0.981	0.004	0.964	0.008	0.963	0.016	0.941	0.012
de-hr-hu-pl	0.961	0.009	0.977	0.005	0.962	0.014	0.971	0.005	0.974	0.009	0.959	0.009	0.965	0.007	0.929	0.023
cs-de-hr	0.961	0.004	0.980	0.005	0.965	0.008	0.972	0.004	0.976	0.003	0.948	0.011	0.967	0.008	0.923	0.006
cs-hr-hu-pl	0.960	0.005	0.981	0.004	0.944	0.018	0.971	0.003	0.980	0.003	0.961	0.007	0.957	0.006	0.927	0.016
de-hr-hu	0.959	0.007	0.975	0.008	0.964	0.011	0.973	0.005	0.978	0.002	0.951	0.005	0.957	0.009	0.922	0.018
de-hu	0.959	0.014	0.979	0.004	0.969	0.008	0.953	0.025	0.981	0.006	0.949	0.011	0.959	0.013	0.947	0.014
cs-de-hu	0.959	0.008	0.975	0.008	0.968	0.010	0.958	0.005	0.976	0.005	0.944	0.008	0.960	0.012	0.938	0.016
hu-pl	0.958	0.008	0.980	0.003	0.943	0.009	0.963	0.010	**0.982**	0.005	0.963	0.003	0.959	0.012	0.940	0.014
de-hr	0.957	0.009	0.977	0.007	0.967	0.008	**0.977**	0.005	0.973	0.004	0.950	0.008	0.965	0.009	0.910	0.025
hr-hu-pl	0.956	0.007	0.977	0.005	0.947	0.012	0.969	0.006	0.979	0.005	0.964	0.006	0.952	0.009	0.912	0.028
cs-hr-hu	0.956	0.016	0.977	0.008	0.942	0.031	0.970	0.008	0.978	0.008	0.946	0.024	0.955	0.017	0.924	0.029
hr-pl	0.955	0.011	0.978	0.005	0.941	0.022	0.976	0.004	0.974	0.006	0.966	0.006	0.957	0.012	0.919	0.034
cs-pl	0.955	0.017	0.977	0.012	0.941	0.024	0.961	0.014	0.973	0.013	0.960	0.017	0.960	0.022	0.935	0.027
pl	0.954	0.021	0.977	0.015	0.925	0.051	0.965	0.014	0.970	0.016	**0.968**	0.003	0.958	0.028	0.945	0.017
cs	0.954	0.015	0.976	0.002	0.941	0.022	0.958	0.018	0.971	0.010	0.946	0.015	0.960	0.016	0.941	0.015
cs-hu	0.953	0.010	0.976	0.004	0.937	0.032	0.961	0.012	0.978	0.005	0.942	0.007	0.957	0.009	0.922	0.014
cs-hr	0.948	0.014	0.975	0.005	0.937	0.034	0.967	0.004	0.974	0.008	0.944	0.014	0.954	0.020	0.902	0.032
hr-hu	0.947	0.020	0.972	0.010	0.937	0.030	0.970	0.007	0.981	0.005	0.941	0.017	0.945	0.026	0.905	0.042
de	0.946	0.022	0.968	0.011	0.970	0.007	0.933	0.033	0.961	0.019	0.932	0.020	0.959	0.014	0.934	0.024
hu	0.934	0.030	0.968	0.013	0.927	0.038	0.941	0.037	0.980	0.006	0.929	0.028	0.934	0.016	0.910	0.027
hr	0.933	0.023	0.971	0.010	0.923	0.037	0.971	0.007	0.965	0.020	0.936	0.017	0.942	0.023	0.859	0.058

Obr. 2 Porovnanie detekčných schopností (priemerná hodnota AUC ROC naprieč základnými modelmi) dotrénovaných detekčných metód pre jednotlivé testovacie jazyky na základe kombinácie trénovacích jazykov [1].

Pre overenie odolnosti detekčných metód voči útokom (s účelom vyhnutia sa detekcii) sme použili homoglyfný (HomoglyphAttack) a parafrázovací útok (pomocou DeepSeek-R1-Distill-Qwen-32B) na 2800 vzoriek testovacej sady (100 textov pre každú triedu, doménu a jazyk). Z dotrénovaných detektorov sme zvolili na vyhodnotenie de-hu-pl kombináciu (reprezentujúca tri vetvy jazykových rodín a dosahujúca vynikajúcu úspešnosť). Výsledky vo forme relatívneho poklesu detekčnej schopnosti sú zobrazené na Obr. 3. Z výsledkov vyplýva, že dotrénované jazykové modely sú narobustnejšie, teda najodolnejšie voči útokom (najmä tie založené na väčších verziách jazykových modelov). Homoglyfný útok

znížil detekčnú schopnosť omnoho viac ako použitý parafrázovací útok, okrem prípadu Detection-Longformer, kde homoglyfný útok dokonca zvýšil úspešnosť detekcie. Detekčné metódy založené na štatistických metrikách sú náchylné na oba typy útokov, pričom homoglyfný útok môže znížiť hodnotu AUC ROC až o 95 % (v prípade Fast-DetectGPT pre nemčinu).

Detector	Subset	All	cs	de	hr	hu	pl	sk	sl
Llama-3.2-3B	paraphrased	**1.1478**	**1.3756**	**0.6972**	**0.8615**	**0.4465**	**0.5965**	**1.4596**	**1.7855**
	homoglyph	-3.9598	-1.6017	-4.4729	-1.5661	-3.2416	-4.6977	-4.3851	-5.1506
mDeBERTa-v3-base	paraphrased	**1.0807**	**1.1490**	**0.8103**	**1.1151**	**0.5387**	**0.7277**	**1.7101**	**1.1182**
	homoglyph	-20.2776	-12.8004	-30.8013	-16.4754	-16.3152	-24.4904	-15.9545	-27.1115
Gemma-2-2B	paraphrased	**1.9915**	**1.2283**	**1.0638**	**2.8481**	**0.2619**	**0.7575**	**1.6004**	**2.7112**
	homoglyph	-6.5255	-4.0138	-3.3260	-5.9583	-2.5900	-6.5343	-6.8872	-9.4358
XLM-RoBERTa-base	paraphrased	**1.3034**	**1.4131**	**0.6518**	**1.5217**	**0.7100**	**0.1951**	**2.6804**	**2.2970**
	homoglyph	-45.3002	-33.6366	-60.0206	-42.7661	-34.3776	-44.8501	-49.0268	-55.7142
Fast-DetectGPT	paraphrased	**-11.5706**	**-4.8167**	**-2.4718**	**-10.1836**	**-35.0241**	**-13.9639**	**-4.1148**	**-8.8522**
	homoglyph	-89.8062	-86.4184	-95.0409	-93.2612	-90.0897	-89.4144	-83.9193	-90.7533
Binoculars	paraphrased	**-8.5022**	**-2.6775**	**-0.7753**	**-7.3183**	**-26.3440**	**-10.0776**	**-2.1252**	**-6.3132**
	homoglyph	-65.0519	-62.6195	-70.4672	-70.1066	-59.7577	-64.1452	-59.3729	-70.8711
LLM-Deviation	paraphrased	**-2.7774**	**-2.0273**	**1.2324**	**-0.9388**	**-17.9738**	**-8.1312**	**0.6255**	**-2.1857**
	homoglyph	-48.1313	-43.5384	-64.7306	-49.6675	-43.9930	-51.0777	-51.2818	-55.4268
BLOOMz-3B-mixed-detector	paraphrased	**25.2145**	**29.8580**	**24.9264**	**21.5639**	**20.4258**	**27.2610**	**25.5008**	**28.2813**
	homoglyph	-37.4430	-33.9972	-37.3355	-38.9864	-38.1557	-36.4922	-35.7036	-43.7688
ChatGPT-detector-RoBERTa-Chinese	paraphrased	**0.2833**	**-3.5264**	**-1.2455**	**4.1424**	**2.4885**	**-1.5669**	**4.5022**	**-7.2485**
	homoglyph	-14.6194	-17.6742	-21.8579	-19.9088	-13.9314	-13.3594	-17.1165	-25.4668
Detection-Longformer	paraphrased	-9.0621	-13.0217	-15.8487	-2.9369	-14.4351	-6.7924	-10.6439	-2.6852
	homoglyph	**10.3109**	**4.7557**	**-2.5400**	**13.2644**	**5.0531**	**6.6735**	**16.3349**	**26.2238**

Obr. 3 Zníženie úspešnosti detekcie (relatívny percentuálny rozdiel v dosiahnutom AUC ROC oproti pôvodným dátam) pri modifikácii testovacej množiny pomocou útočných metód [1].

Výsledky tejto časti výskumu boli publikované vo forme preprintu [1], momentálne sú v štádiu posudzovania v rámci ACL Rolling Review a budú publikované na niektorej z top *ACL konferencií (príp. workshopov).

2.2 Zvýšenie robustnosti MGT detekcie

Ako bolo ukázané v predchádzajúcej štúdii [1] opísanej vyššie, MGT detektory sú zraniteľné voči útokom ako aj voči textom, ktorých distribúcia tokenov sa odlišuje od dát, na ktorých boli trénované (napr. odlišné jazyky alebo domény, príp. nové generátory alebo aj len odlišné nastavenia generovania textov). Preto sme v rámci ďalšej časti výskumu [2] navrhli spôsob kompozície trénovacej sady textov, predspracovania textov, ako aj voľbu škálovateľnej metódy dotrénovania detekčných modelov založených na LLM. Výsledkom sú robustnejšie detektory ako z hľadiska odolnosti voči útokom, tak aj z hľadiska dát mimo trénovacej distribúcie.

Predchádzajúci výskum ukázal dobrú medzijazykovú transferabilitu detekčných schopností pri trénovaní na väčšom množstve jazykov. Preto sme zaradili do dotrénovania detektorov veľkú množinu jazykov (vyše 44). Kvôli robustnosti voči novým doménam sme (podobne ako v predchádzajúcej štúdii) skombinovali dve domény (novinové články a sociálne médiá) pre rovnaké páry generátorov a jazykov. Na zvýšenie odolnosti voči útokom sme do trénovania zahrnuli vzorky modifikované pomocou špecializovaného útoku DFT-Fooler, parafrázovaných pomocou ChatGPT a spätne preložených pomocou m2m100-1.2B. Tieto modifikované dáta slúžia na augmentáciu trénovacích dát. Okrem toho sme navrhli kombinovaný útok zvaný HomoglyphJoinerAttack, ktorý kombinuje pseudonáhodný homoglyfný útok s vkladaním neviditeľného znaku. Tento útok je použitý v rámci predspracovania textov na modifikáciu časti MGT vzoriek. Predspracovanie ako trénovacích, tak aj testovacích vzoriek zahŕňa taktiež anonymizáciu a konverziu na malé písmená. Na dotrénovanie je použitá 4-bitová kvantizácia v rámci QLoRA techniky. Ako validačná množina sú použité rôznorodé dáta z MIX2k datasetu (7 jazykov, 50 generátorov a rozličné domény).

Okrem AUC ROC sme na evaluáciu použili aj metriku TPR@5%FPR, ktorá odzrkadluje množstvo správne detegovaných MGT vzoriek pri stanovenej maximálnej akceptovateľnej miere (5 %) nesprávne identifikovaných ľudských textov. Výsledky na Obr. 4 ukazujú, že navrhnutá metóda robustného dotrénovania detekčných modelov čiastočne znížila úspešnosť detekcie na dátach v rámci distribúcie, avšak výrazne zvýšila úspešnosť detekcie na dátach mimo trénovacej distribúcie, t.j. zvýšila robustnosť. Navyše výpočtovo efektívny spôsob dotrénovania pomocou kvantizácie a minimalizácie počtu trénovaných parametrov (QLoRA) umožnil škálovať veľkosť použitého modelu. To prinieslo najúspešnejší multilingválny MGT detektor založený na modeli Gemma-2-9b-it. Ale aj menej výpočtovo náročný model Qwen2-1.5B umožňujúci nasadanie detekčnej služby bez GPU akcelerácie.

Detector	MULTITuDE_v3		MultiSocial	
	AUC ROC	TPR@5%FPR	AUC ROC	TPR@5%FPR
Gemma-2-9b-it	0.9914	0.9798	0.9563	0.8400
Qwen2-0.5B	0.9785	0.9316	0.9582	0.8413
Qwen2-1.5B	0.9883	0.9460	0.9549	0.7781
Yi-1.5-6B	0.9748	0.9115	0.9474	0.8008
mDeBERTa-v3-base	**0.9959**	0.9797	0.9540	0.7750
mDeBERTa (baseline)	0.9944	**0.9875**	**0.9746**	**0.8862**

Detector	MIX		SemEval	
	AUC ROC	TPR@5%FPR	AUC ROC	TPR@5%FPR
Gemma-2-9b-it	**0.8901**	**0.5227**	**0.9448**	0.8284
Qwen2-0.5B	0.6499	0.0000	0.8434	0.0000
Qwen2-1.5B	0.7588	0.3064	0.9391	**0.8287**
Yi-1.5-6B	0.8167	0.0000	0.8922	0.0000
mDeBERTa-v3-base	0.6669	0.0502	0.8666	0.6843
mDeBERTa (baseline)	0.5502	0.0000	0.8305	0.0000

Obr. 4 Porovnanie detekčných schopností robustne dotrénovaných modelov v porovnaní s klasickou metódou dotrénovania (baseline). MULTITuDE_v3 a MultiSocial predstavujú datasety obsahujúce dáta v rámci trénovacej distribúcie, pričom MIX a SemEval datasety obsahujú dáta mimo distribúcie [3].

Okrem overenia robustnosti vzhľadom na dáta mimo trénovacej distribúcie sme overili aj odolnosť voči útokom. Pre tento účel sme použili náš dataset MULTITuDE_v2. Z 10 obsiahnutých útočných metód sme zvolili 8, ktoré sú použiteľné v multilingválnom prostredí. Z výsledkov na Obr. 5 vyplýva, že modely dotrénované navrhnutým robustným spôsobom sú imúnne voči týmto útokom (aj voči tým, ktoré neboli zahrnuté do dotrénovania). Pôvodný (baseline) detektor utrpel vplyvom útoku HomoglyphAttack zníženie hodnoty AUC ROC o 27 % a menšie zníženie aj vplyvom ostatných útokov. To potvrdzuje všeobecne vyššiu robustnosť navrhnutých MGT detektorov.

Obfuscator	Gemma-2-9b-it	Qwen2-1.5B	mDeBERTa-v3-base	baseline
original	0.9757	0.9719	**0.9887**	0.9821
adversarial attacks				
ALISON	0.9790	0.9744	**0.9872**	0.9804
DFTFooler	0.9834	0.9783	**0.9946**	0.9776
backtranslation				
m2m100-1.2B	0.9842	0.9840	**0.9918**	0.9780
nllb-200-distilled-1.3B	0.9843	0.9878	**0.9913**	0.9819
paraphrasing				
ChatGPT	0.9816	0.9905	**0.9953**	0.9926
text edits				
HomoglyphAttack	0.9860	0.9850	**0.9986**	0.7213
GPTZeroBypasser	0.9880	**1.0000**	0.9987	0.8505
GPTZzzs	0.9806	0.9776	**0.9917**	0.9793

Obr. 5 Vplyv útočných techník na MGT detekčnú schopnosť (AUC ROC) navrhnutých robustných detektorov v porovnaní s baseline detektorom [3].

Výsledky tejto časti výskumu boli publikované vo forme preprintu [3], momentálne sú v štádiu posudzovania v rámci ACL Rolling Review a budú publikované na niektorej z top *ACL konferencií (príp. workshopov). Táto metóda robustnej detekcie bola navyše modifikovaná v rámci zdieľanej úlohy Voight-Kampff Generative AI Detection 2025 v rámci PAN labu konferencie CLEF 2025. Naše riešenie mdok [4] obsadilo 1. miesto v obidvoch podúlohách, pričom prvá bola zameraná na robustnú MGT detekciu (aj odolnosť voči útokom) anglických textov a druhá bola zameraný na viactriednu klasifikáciu 6 typov hybridných textov (rôzne stupne kolaborácie medzi AI a človekom).

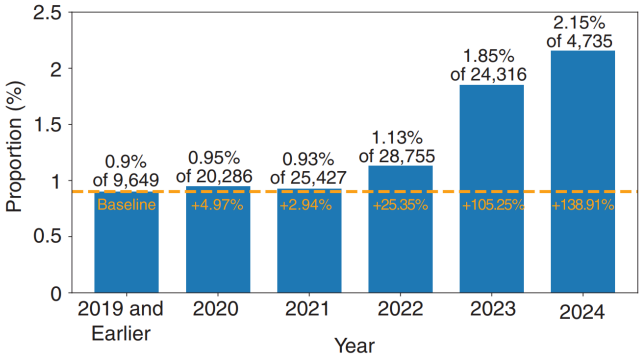
2.3 Aplikácia robustnej MGT detekcie

Výsledky predchádzajúcej štúdie ohľadom robustnej detekcie MGT [3] sme použili pri aplikácii detekcie MGT ako indikátora dezinformačného obsahu. Pre tento účel sme použili dataset MultiClaim, ktorý obsahuje príspevky z rôznych sociálnych médií, ktorých tvrdenia sú namapované na reporty profesionálov overujúcich fakty (ang. fact-checker). Na základe týchto overení predstavuje drvivá väčšina MultiClaim textov dezinformácie.

V tejto časti výskumu sme použili náš robustný multilingválny detektor založený na dotrénovanom Gemma-2-9b-it modeli [3] (označený MultiDomain). Pre účely aplikácie na neoznačkovaných dátach (bez tzv. ground truth) bolo potrebné minimalizovať mieru FPR (angl. false positive rate), teda nesprávne identifikovaných ľudských textov. Preto sme tento model taktiež dotrénovali na GenAI datasete z nedávnej viacjazyčnej zdieľanej úlohy. Tieto dva modely sme použili v kombinovanej detekcii, pričom ako MGT boli označené texty, pri ktorých si bol aspoň jeden z detektorov úplne istý (pravdepodobnosť 1,0) a druhý si nebol

úplne istý negatívnou predikciou (pravdepodobnosť 0,0). Navyše sme niektoré jazyky z GenAI detektora vylúčili kvôli vysokej miere FPR na existujúcich označkovaných datasetoch (arabčina, nemčina, ruština a taliančina), t.j. v týchto jazykoch sa zohľadnila len predikcia MultiDomain detektora. Pri overené na 5 existujúcich multilingválnych datasetoch dosiahla takáto kombinovaná detekcia maximálne 5,7 % miery FPR, takže detegované MGT v MultiClaim dátach sú s veľkou pravdepodobnosťou skutočne generované.

Na Obr. 6 sú zobrazené výsledky detekcie vo forme pomeru (proporcie) MGT z MultiClaim dát pre jednotlivé roky. Rok 2019 a skôr považujeme za baseline, keďže vtedy ešte jazykové modely neboli dostatočne úspešné pri generovaní multilingválnych textov. Odhadnuté množstvo približne 0,9 % textov teda môžeme považovať za falošne pozitívne predikcie, resp. MGT vygenerované strojmi ešte pred rozkvetom LLM (napr. prekladačmi, korektormi). Výsledky preukazujú, že množstvo MGT v overovaných textoch zo sociálnych médií každým rokom narastá, pričom najväčší nárast je pozorovaný medzi rokmi 2022 a 2023, čo korešponduje so zverejnením ChatGPT v novembri 2022.



Obr. 6 Pomer textov detegovaných ako MGT v datasete MultiClaim agregovaných do príslušných rokov [5].

Výsledky tejto časti výskumu boli publikované v magazíne Computer [5], vydávanom IEEE Computer Society.

V ďalšej časti výskumu bola naša viactriedna robustná klasifikácia vo forme mdok [4] ďalej modifikovaná na detekciu konkrétneho generátora MGT vo viacerých jazykoch (angl. multilingual authorship attribution). Okrem nášho mdok riešenia sme prispôsobili tejto úlohe aj ďalšie existujúce binárne detektory MGT a vyhodnotili na datasete MULTITuDE_v3 s vyváženým množstvom vzoriek pre každý jazyk a generátor. Ako evaluačnú metriku sme použili Macro F1, ktorá priemeruje úspešnosť detekcie pre každú triedu. Na Obr. 7 sú zobrazené výsledky pre každý testovací jazyk. V tomto prípade boli na trénovanie použité

vzorky zo všetkých jazykov. Ako možno vidieť, nie len mdok detektor ale aj OTBDetector a Qwen3-4B-Base detektor poskytujú výbornú schopnosť detekcie vo všetkých jazykoch, pričom čínština dosiahla najnižšiu úspešnosť. Metódy založené na štatistických metrikách sú pre túto úlohu nepoužiteľné vzhľadom na nízku úspešnosť detekcie.

Lang. family →	Germanic			Romance			Slavic-Latin					Slavic-Cyrillic			Others				
Method ↓	de	en	nl	es	pt	ro	cs	hr	pl	sk	sl	bg	ru	uk	hu	el	ar	zh	all
Qwen3-4B-Base	**0.92**	0.91	**0.95**	**0.92**	**0.93**	**0.95**	**0.96**	0.95	0.94	0.97	0.95	**0.95**	**0.92**	0.93	**0.93**	0.93	**0.96**	0.85	**0.93**
mdok	0.92	**0.91**	0.95	0.91	0.93	0.94	0.95	**0.96**	**0.94**	**0.97**	**0.95**	0.93	0.91	**0.93**	0.93	**0.94**	0.96	**0.87**	0.93
OTBDetector	0.87	0.78	0.91	0.85	0.89	0.93	0.93	0.93	0.92	0.96	0.94	0.93	0.87	0.91	0.91	0.92	0.95	0.80	0.90
XLM-R-large	0.81	0.65	0.84	0.76	0.80	0.87	0.88	0.88	0.88	0.93	0.90	0.87	0.78	0.84	0.86	0.88	0.90	0.72	0.84
RoBERTa-large	0.78	0.72	0.81	0.74	0.80	0.84	0.83	0.83	0.81	0.85	0.84	0.63	0.63	0.67	0.76	0.59	0.70	0.60	0.75
StatEnsemble	0.49	0.33	0.55	0.45	0.47	0.48	0.43	0.43	0.50	0.43	0.31	0.51	0.48	0.48	0.50	0.41	0.40	0.35	0.45
Fast-DetectGPT	0.25	0.12	0.25	0.18	0.20	0.19	0.23	0.22	0.26	0.18	0.20	0.31	0.31	0.31	0.30	0.16	0.17	0.16	0.23
Binoculars	0.20	0.15	0.22	0.15	0.18	0.24	0.14	0.14	0.23	0.13	0.17	0.07	0.13	0.08	0.13	0.14	0.12	0.14	0.16
Average	0.65	0.57	0.68	0.62	0.65	0.68	0.67	0.67	0.69	0.68	0.66	0.65	0.63	0.64	0.66	0.62	0.64	0.56	0.65
Writing script →	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Lat	Cyr	Cyr	Cyr	Lat	Grk	Arab	Han	

Obr. 7 Porovnanie úspešnosti detekcie generátora MGT (Macro F1) [2].
Označenie písma: Lat = Latin, Cyr = Cyrillic, Grk = Greek, Arab = Arabic, Han = Hanzi.

Výsledky tejto časti výskumu boli publikované vo forme preprintu [2], momentálne sú v štádiu posudzovania v rámci ACL Rolling Review a budú publikované na niektorej z top *ACL konferencií (príp. workshopov).

3 Záver

Náš výskum v oblasti detekcie strojovo-generovaných textov v stredoeurópskych jazykoch identifikoval určité (ale nie veľmi výrazné) rozdiely v detekčných schopnostiach dotrénovaných modelov v závislosti na kombinácii trénovacích jazykov. Najlepšie výsledky dosahovali detektory dotrénované aspoň na troch rozličných jazykoch. Dotrénované špecializované modely dosahovali výrazne lepšie výsledky ako existujúce všeobecné metódy založené na štatistických metrikách alebo existujúce predtrénované detektory. Rovnako ukázali aj vyššiu robustnosť v zmysle odolnosti voči homoglyfným a parafrázovacím útokom ako existujúce detektory. Navyše výskum v oblasti zvýšenia robustnosti detektorov ukázal, že detektory dotrénované navrhnutou robustnou metódou (zahŕňajúcou aj niektoré útočné vzorky do trénovania) majú vyššiu odolnosť voči zmene distribúcie dát (voči trénovacej sade, napr. odlišné generátory textov), ako aj imunitu voči testovaným útokom. Takéto robustné detektory majú aj praktické využitie v oblasti boja proti dezinformáciám, kde boli použité napr. na identifikáciu množstva strojovo-generovaných textov v overovaných príspevkoch zo sociálnych médií.

4 Referencie

[1] Dominik Macko and Jakub Kopal. 2025. CEAID: Benchmark of Multilingual Machine-Generated Text Detection Methods for Central European Languages. arXiv preprint arXiv:2509.26051.

[2] Lucio La Cava, Dominik Macko, Róbert Móro, Ivan Srba, and Andrea Tagarelli. 2025. Authorship Attribution in Multilingual Machine-Generated Texts. arXiv preprint arXiv:2508.01656.

[3] Dominik Macko, Robert Moro, Ivan Srba. 2025. Increasing the Robustness of the Fine-tuned Multilingual Machine-Generated Text Detectors. arXiv preprint arXiv:2503.15128.

[4] Dominik Macko. 2025. mdok of KInIT: Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection. In *Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

[5] Dominik Macko, Aashish Anantha Ramakrishnan, Jason Samuel Lucas, Robert Moro, Ivan Srba, Adaku Uchendu, and Dongwon Lee. 2026. Beyond speculation: Measuring the growing presence of LLM-generated texts in multilingual disinformation. In *Computer*, vol. 59, no. 2. IEEE. DOI: 10.1109/MC.2025.3592765.