V3.2 Výskumná správa o optimalizovanej architektúre systému na detekciu 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

Obsah

1 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 prezentuje výsledky výskumu zameraného na optimalizáciu architektúry systému pre detekciu strojovo-generovaného textu v stredoeurópskom multilingválnom prostredí (vychádzame z definície podľa Bideleux and Jeffries (2007) zahŕňajúcu češtinu, chorvátčinu, maďarčinu, nemčinu, poľštinu, slovenčinu a slovinčinu). Správa nadväzuje na predchádzajúce aktivity projektu RobIndAI, predovšetkým na výskum metód detekcie strojovo-generovaného textu, vytvorenie/rozšírenie datasetov o stredoeurópske jazyky a výskum robustnosti detekčných metód. V predchádzajúcej časti výskumu opísanej vo výstupe V3.1 sme identifikovali dobrú schopnosť medzijazykovej prenositeľnosti detekčnej schopnosti medzi jazykmi stredoeurópskeho regiónu. Preto sa v tejto výskumnej správe venujeme porovnaniu architektúry detekčného systému pozostávajúcej zo samostatných detektorov špecializovaných pre každý jazyk osobitne voči architektúre systému s univerzálnym detektorom trénovaným na kombinácii jazykov. Berieme do úvahy aj výslednú robustnosť ako aj výpočtovú náročnosť potenciálne nasadeného systému.

Táto výskumná správa je rozdelená na tri časti. V prvej časti (Kap. 2.1) sa zameriavame na porovnávaciu štúdiu architektúr systémov MGT detekcie vzhľadom na úspešnosť detekcie pre jazyky stredoeurópskeho regiónu. V druhej časti (Kap. 2.2) sa zameriavame na porovnanie robustnosti týchto architektúr MGT systémov voči neznámym jazykom, príp. pri potrebe jemnejšej granularity pri identifikácii generatívneho modelu. Tretia časť (Kap. 2.3) je zameraná na porovnanie týchto architektúr vzhľadom na odhadnutú výpočtovú náročnosť.

2 Porovnanie architektúr MGT detekčných systémov

V našom výskume sa zameriavame na porovnanie rôznych architektúr MGT detekčných systémov, kombinujúcich metódy založené na jazykových AI modeloch. Cieľom je identifikácia optimálnej architektúry detekčného systému pre stredoeurópske jazyky s dôrazom na viaceré požiadavky, ako napr. úspešnosť detekcie, ale aj výpočtová efektivita systému.

2.1 Úspešnosť detekcie MGT v stredoeurópskych jazykoch

Pre účel porovnania detekčných schopnosti rôznych architektúr kombinovaných detekčných systémov použijeme dataset CEAID predstavený vo výskumnej správe V3.1 a opísaný vo výskumnom článku [1], ktorý je prijatý do Findings časti konferencie ACL 2026 (A* core rank). CEAID dataset je v súčasnosti zverejnený na platforme Zenodo[1] pre nekomerčné výskumné účely.

Tento dataset bol navrhnutý s cieľom zabezpečiť konzistentné medzijazykové porovnanie metód detekcie MGT v jazykoch stredoeurópskeho regiónu. Dataset vznikol kombináciou dvoch existujúcich datasetov MULTITuDE_v3 (novinové články) a MultiSocial (texty zo sociálnych médií), teda pokrýva dve textové domény. Výber dát bol realizovaný tak, aby každý jazyk obsahoval minimálne 200 testovacích vzoriek pre každú triedu (ľudský vs. MGT) a pre každú doménu. Výsledný dataset zahŕňa sedem jazykov stredoeurópskeho regiónu: češtinu, chorvátčinu, maďarčinu, nemčinu, poľštinu, slovenčinu a slovinčinu. Slovenčina a slovinčina neboli použité pri trénovaní modelov z dôvodu nedostatočného množstva dát zo sociálnych médií, no boli zahrnuté do testovacej časti datasetu. MGT boli vytvorené pomocou ôsmich LLM. Šesť modelov bolo spoločných pre obe domény: Aya-101, GPT-3.5-Turbo-0125, Mistral-7B-Instruct-v0.2, OPT-IML-Max-30B, v5-Eagle-7B-HF a Vicuna-13B. Okrem nich bol v novinových článkoch použitý model Llama-2-70B-chat-hf a v doméne sociálnych médií model Gemini. Celkovo dataset obsahuje 132 168 trénovacích a 55 930 testovacích vzoriek naprieč všetkými jazykmi a doménami. Najväčšie zastúpenie má nemčina s viac ako 28 tisíc trénovacími vzorkami, zatiaľ čo slovenčina a slovinčina obsahujú iba tréningové dáta z domény novinových článkov.

[1] https://doi.org/10.5281/zenodo.20391070

Kvôli zabezpečeniu férového porovnania systémov boli testovacie dáta pseudonáhodne zredukované na 250 vzoriek pre každú kombináciu triedy, domény a jazyka (teda perfektné vyvážanie dát). Trénovacie dáta na dotrénovanie používali vždy rovnaké množstvo vzoriek. Na evaluáciu sme zvolili metriku AUC ROC, ktorá reprezentuje všeobecnú detekčnú schopnosť (bez stanovenia konkrétneho prahu pre klasifikáciu medzi dvoma triedami). Jazyky sú označené dvojpísmenovým kódom podľa ISO 639-1.

Pre účely porovnania architektúr detekčných systémov sme sa zamerali jazykové modely dotrénované na tomto datasete na úlohu binárnej klasifikácie (teda detekcie MGT). Zvolené jazykové modely zahŕňajú rôzne architektúry (enkóder vs. dekóder), rodiny a veľkosti: mDeBERTa-v3-base, XLM-RoBERTa-base, Llama-3.2-3B, Gemma-2-2B.

Porovnávané architektúry detekčného systému:

a) Jeden viacjazyčne dotrénovaný model (single) – dotrénovanie na všetkých piatich trénovacích jazykoch kombinovane.
b) Kombinácia jednojazyčne dotrénovaných modelov pomocou priemeru pravdepodobnosti predikcie (mean) – všetkých päť jazykovo-špecifických modelov vykoná predikciu a ich výsledné pravdepodobnosti sa spriemerujú.
c) Kombinácia jednojazyčne dotrénovaných modelov pomocou preferencie detektora korešpondujúceho jazyka (perlang) – predikciu vykoná len model, ktorý bol dotrénovaný v danom jazyku. V prípade neznámeho jazyka vykonajú predikciu všetky a ich pravdepodobnosti sa spriemerujú.

Výsledky na Obr. 1 naznačujú, že alternatíva b) (označená mean) dosahuje najlepšie výsledky v prípade mDeBERTa a Llama modelov, ale najhoršie v prípade XLM-RoBERTa a Gemma. Alternatíva a) dosiahla pri prvých dvoch modeloch najhoršie výsledky a pri druhých dvoch modeloch naopak najlepšie. Takéto výsledky sú teda nejednoznačné, spôsobené veľmi podobnou schopnosťou detekcie jednotlivých systémov (rozdiel v AUC ROC hodnotách do 1 % pre daný základný model). Keďže výsledky môžu byť ovplyvnené agregovaným testovaním na oboch doménach, overili sme porovnanie detekčných schopností aj pre každú doménu samostatne (Obr. 2 a Obr. 3).

System	All	cs	de	hr	hu	pl	sk	sl
mdeberta-v3-base_mean	**0.9793**	0.9878	**0.9766**	0.9816	**0.9883**	**0.9729**	0.9775	**0.9749**
Llama-3.2-3B_mean	0.9768	**0.9883**	0.9750	**0.9838**	0.9864	0.9675	**0.9802**	0.9564
mdeberta-v3-base_perlang	0.9753	0.9787	0.9751	0.9799	0.9856	0.9720	0.9775	**0.9749**
mdeberta-v3-base_single	0.9730	0.9822	0.9690	0.9771	0.9830	0.9705	0.9716	0.9581
Llama-3.2-3B_perlang	0.9699	0.9761	0.9690	0.9721	0.9775	0.9683	**0.9802**	0.9564
Llama-3.2-3B_single	0.9674	0.9842	0.9645	0.9743	0.9807	0.9609	0.9727	0.9330
gemma-2-2b_single	0.9647	0.9835	0.9693	0.9730	0.9762	0.9614	0.9697	0.9238
gemma-2-2b_perlang	0.9639	0.9755	0.9764	0.9639	0.9827	0.9698	0.9733	0.9405
xlm-roberta-base_single	0.9621	0.9778	0.9484	0.9744	0.9748	0.9541	0.9606	0.9497
xlm-roberta-base_perlang	0.9594	0.9756	0.9608	0.9666	0.9729	0.9638	0.9501	0.9317
xlm-roberta-base_mean	0.9553	0.9740	0.9349	0.9714	0.9753	0.9423	0.9501	0.9317
gemma-2-2b_mean	0.9534	0.9880	0.9509	0.9690	0.9761	0.9604	0.9733	0.9405

Obr. 1 Porovnanie detekčných schopností (AUC ROC) zvolených detekčných systémov pre jednotlivé testovacie jazyky v oboch doménach kombinovane.

System	All	cs	de	hr	hu	pl	sk	sl
Llama-3.2-3B_mean	**0.9956**	0.9971	0.9918	0.9985	**0.9986**	**0.9945**	0.9952	0.9958
mdeberta-v3-base_mean	0.9954	**0.9991**	**0.9952**	0.9971	0.9956	0.9916	0.9960	**0.9970**
mdeberta-v3-base_perlang	0.9945	0.9986	0.9950	0.9965	0.9938	0.9934	0.9960	**0.9970**
mdeberta-v3-base_single	0.9939	0.9984	0.9946	0.9963	0.9929	0.9910	0.9948	0.9929
Llama-3.2-3B_perlang	0.9934	0.9964	0.9853	**0.9996**	0.9954	0.9911	0.9952	0.9958
Llama-3.2-3B_single	0.9919	0.9982	0.9838	0.9948	0.9921	0.9892	0.9950	0.9910
gemma-2-2b_perlang	0.9906	0.9962	0.9944	0.9946	0.9921	0.9925	0.9966	0.9928
gemma-2-2b_single	0.9904	0.9982	0.9909	0.9920	0.9831	0.9866	**0.9974**	0.9878
xlm-roberta-base_single	0.9824	0.9896	0.9728	0.9876	0.9759	0.9847	0.9769	0.9910
xlm-roberta-base_perlang	0.9807	0.9935	0.9785	0.9928	0.9856	0.9918	0.9503	0.9828
gemma-2-2b_mean	0.9798	0.9972	0.9875	0.9937	0.9861	0.9877	0.9966	0.9928
xlm-roberta-base_mean	0.9734	0.9877	0.9604	0.9866	0.9755	0.9780	0.9503	0.9828

Obr. 2 Porovnanie detekčných schopností (AUC ROC) zvolených detekčných systémov pre jednotlivé testovacie jazyky v doméne novinových článkov.

System	All	cs	de	hr	hu	pl	sk	sl
mdeberta-v3-base_mean	**0.9527**	0.9659	0.9438	0.9614	**0.9798**	**0.9472**	0.9503	**0.9328**
Llama-3.2-3B_mean	0.9493	**0.9777**	0.9413	**0.9618**	0.9733	0.9297	**0.9612**	0.8977
mdeberta-v3-base_perlang	0.9441	0.9455	**0.9515**	0.9496	0.9732	0.9381	0.9503	**0.9328**
xlm-roberta-base_mean	0.9412	0.9639	0.9218	0.9600	0.9763	0.9170	0.9504	0.8842
mdeberta-v3-base_single	0.9380	0.9514	0.9246	0.9447	0.9743	0.9376	0.9384	0.8971
Llama-3.2-3B_perlang	0.9378	0.9503	0.9400	0.9364	0.9562	0.9361	**0.9612**	0.8977
Llama-3.2-3B_single	0.9372	0.9664	0.9315	0.9486	0.9689	0.9235	0.9489	0.8691
xlm-roberta-base_single	0.9365	0.9633	0.9146	0.9572	0.9732	0.9136	0.9432	0.8875
xlm-roberta-base_perlang	0.9347	0.9522	0.9383	0.9338	0.9618	0.9240	0.9504	0.8842
gemma-2-2b_single	0.9313	0.9631	0.9334	0.9468	0.9686	0.9240	0.9324	0.8483
gemma-2-2b_perlang	0.9286	0.9486	0.9460	0.9228	0.9703	0.9335	0.9405	0.8525
gemma-2-2b_mean	0.9220	0.9741	0.9221	0.9347	0.9611	0.9257	0.9405	0.8525

Obr. 3 Porovnanie detekčných schopností (AUC ROC) zvolených detekčných systémov pre jednotlivé testovacie jazyky v doméne sociálnych médií.

Pri novinových článkoch aj keď došlo k istej zmene celkového poradia najlepších modelov podľa AUC ROC, k zmene poradia v rámci alternatív pre jednotlivé základné modely nedošlo. V doméne sociálnych médií však vidíme, že alternatíva b) dosiahla najvyššie skóre aj v prípade XML-RoBERTa modelu. V prípade Gemma modelu sa poradie nezmenilo. Keďže pri troch zo štyroch testovaných základných modelov dosiahla alternatíva b) najlepšie výsledky, hodnotíme ju z tohto hľadiska ako najvhodnejšiu.

Záverom tohto porovnania je, že kombinácia jednojazyčne dotrénovaných modelov pomocou priemeru pravdepodobností jednotlivých predikcií prekonáva jeden multijazyčne dotrénovaný model. Takýto systém si však vyžaduje dostupnosť dostatočného množstva dát v každom jazyku. Prístup k pravdepodobnostiam predikcie tiež nie je vždy možný, príp. takáto naivná kombinácia nie je najvhodnejšou alternatívou (napr. pri viactriednej klasifikácii). Prenositeľnosť do iných jazykov, mimo regiónu stredoeurópskeho priestoru je tiež otázna, vzhľadom na slabšie príbuzenstvo jazykov.

2.2 Robustnosť pri viactriednej klasifikácii a prenositeľnosť do neznámych jazykov

Pre účely porovnania rôzynch architektúr detekčných systémov pre viactriednu klasifikáciu vychádzame zo štúdie [2], čiastočne predstavenej vo výskumnej správe V3.1, ktorá bola priamo zameraná na atribúciu generátorov pri detekcii MGT v 18 jazykoch. Výskumný článok reportujúci túto štúdiu je prijatý do hlavnej časti konferencie ACL 2026 (A* core rank).

Dataset použitý na vyhodnotenie v rámci tohto výskumu je podmnožinou nášho datasetu MULTITuDE_v3, ktorý, ako bolo spomenuté v predchádzajúcej časti, je zameraný na textovú doménu novinových článkov. Z dostupných 21 jazykov sme použili 18 (vymenované na Obr. 4), ktoré obsahujú plne vyváženú množinu kombinácií generátorov a jazykov, pričom sme ako prah zvolili dostupnosť 95 % vzoriek z nášho cieľa 1000 trénovacích vzoriek a 300 testovacích pre každú triedu (generátor). Dáta boli vygenerované siedmimi LLM: Mistral-7B-Instruct-v0.2, OPT-IML-Max-30B, v5-Eagle-7B-HF, Vicuna-13B, Llama-2-70B-Chat-HF, Aya-101 a GPT-3.5-Turbo-0125. Ôsmu triedu predstavujú ľudské texty. Ako je zobrazné na Obr. 4, každý z 18 jazykov obsahuje približne 7 950 trénovacích vzoriek a 2 380 testovacích vzoriek. Tieto jazyky pokrývajú všetky jazyky sledovaného stredoeurópskeho regiónu.

Family	Language	Code	Train	Test
Germanic	Dutch	nl	7958	2386
	English	en	7954	2384
	German	de	7951	2388
Hellenic	Greek	el	7944	2384
Semitic	Arabic	ar	7975	2392
Sino-Tibetan	Chinese	zh	7926	2383
Slavic-Cyrillic	Bulgarian	bg	7954	2386
	Ukrainian	uk	7939	2385
	Russian	ru	7945	2382
Slavic-Latin	Croatian	hr	7951	2384
	Czech	cs	7962	2389
	Polish	pl	7946	2383
	Slovak	sk	7946	2385
	Slovenian	sl	7947	2386
Romanic	Portuguese	pt	7956	2388
	Romanian	ro	7949	2386
	Spanish	es	7947	2387
Uralic	Hungarian	hu	7964	2385
Total	–	–	**143,114**	**42,943**

Obr. 4 Počet vzoriek pre jednotlivé jazyky v trénovacej a testovacej časti zvolených dát z datasetu MULTITuDE_v3 [2].

Ako sledované detekčné metódy, ktorých kombináciu do detekčného systému porovnávame, sme zvolili dotrénované LLM detektory mdok, Qwen3-4B-Base a XML-RoBERTa-Large. Okrem nich sa zameriavame aj na detekčný ensemble pomocou fúzie deviatich štatistických metrík založenej na logistickej regresii (nazvaný StatEnsembleLR). Tento ensemble umožňuje širokospektrálnu analýzu pomocou post-hoc tréningov rôznych klasifikátorov (rôznej kombinácie jazykov) na základe už extrahovaných štatistických metrík, slúžiacich ako vstupné črty. Kvôli výpočtovej náročnosti pre účely tejto analýzy nedotrénujeme nové LLM detektory, ale využijeme už natrénované jednojazyčné detektory použité v štúdii [2] na účel analýzy vplyvu slovanských jazykov na medzijazykovú prenositeľnosť. Teda LLM detektory pokrývajú dotrénovanie na 5 jazykoch stredoeurópskeho regiónu a StatEnsembleLR pokrýva natrénované detektory na všetkých 7 jazykoch. Kvôli viactriednej klasifikácii sme zvolili metriku Macro F1, ktorá priemeruje detekčnú schopnosť naprieč všetkými triedami (generátormi).

Porovnávané architektúry detekčného systému sú podobné ako v predchádzajúcej časti, priemer pravdepodobností predikcie bol však nahradený väčšinovou voľbou kvôli úlohe viactriednej klasifikácie:

a) Jeden viacjazyčne dotrénovaný model (single) – dotrénovanie na všetkých siedmich trénovacích jazykoch kombinovane (dostupný len pre StatEnsembleLR).
b) Kombinácia jednojazyčne dotrénovaných modelov pomocou väčšinovej voľby predikcie (mean) – všetkých päť (resp. sedem v prípade StatEnsembleLR) jazykovo-špecifických modelov vykoná predikciu a pomocou väčšinovej voľby sa vyberie výsledok.
c) Kombinácia jednojazyčne dotrénovaných modelov pomocou preferencie detektora korešpondujúceho jazyka (perlang) – predikciu vykoná len model, ktorý bol dotrénovaný v danom jazyku. V prípade neznámeho jazyka vykonajú predikciu všetky a väčšinová voľba určí výsledok.

Pre porovnanie výsledkov s predchádzajúcou časťou uvádzame najskôr vyhodnotenie binárnej klasifikácie (čiže všetky predikcie iné ako „human“ predstavujú triedu „machine“). Výsledky na Obr. 5 zobrazujú Macro F1 hodnoty pre jednotlivé testovacie jazyky (stĺpce), pričom *All-CE* predstavuje texty všetkých 7 jazykov stredoeurópskeho regiónu a *All* predstavuje všetkých 18 testovacích jazykov. Architektúry detekčného systému sú uvedené v riadkoch, pričom údaje sú dostupné pre všetky 3 dotrénované detektory. Výsledky indikujú, že alternatíva c), označená perlang, dosahuje lepšiu detekčnú schopnosť v jazykoch stredoeurópskeho regiónu ako aj prenositeľnosť do ostatných jazykov datasetu ako alternatíva b), označená mean, ktorá bola vyhodnotená ako najlepšia v predchádzajúcej časti.

System		ar	bg	cs	de	el	en	es	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	All-CE	All
mean	Qwen3-4B-Base	0.47	0.73	0.99	**0.93**	0.56	0.83	0.87	0.98	0.88	0.75	0.95	0.93	0.97	0.52	0.99	0.98	0.70	0.47	0.96	0.84
	XLM-R-large	0.69	0.87	0.96	0.91	0.71	0.74	0.85	0.93	0.91	0.74	0.93	0.73	0.90	0.85	0.98	0.94	0.80	0.62	0.94	0.85
	mdok	0.77	0.93	0.98	0.85	0.78	0.54	0.75	0.98	0.95	0.62	0.96	0.76	**0.97**	0.85	0.98	0.98	0.88	0.47	0.96	0.86
perlang	Qwen3-4B-Base	0.47	0.73	**0.99**	**0.93**	0.56	0.83	0.87	0.98	0.88	0.75	0.97	0.93	0.97	0.52	**1.00**	0.99	0.70	0.47	0.97	0.85
	XLM-R-large	0.69	0.87	0.97	0.91	0.71	0.74	0.85	0.96	0.91	0.74	0.96	0.73	0.90	0.85	0.99	0.97	0.80	0.62	0.95	0.85
	mdok	0.77	0.93	0.99	0.85	0.78	0.54	0.75	**0.99**	0.95	0.62	**0.98**	0.76	**0.97**	0.85	0.99	**0.99**	0.88	0.47	**0.97**	0.87
cs	Qwen3-4B-Base	0.47	0.50	**0.99**	0.91	0.49	0.76	0.84	0.96	0.82	0.56	0.94	0.86	0.95	0.48	0.97	0.96	0.60	0.47	0.94	0.80
	XLM-R-large	0.71	0.84	0.97	0.91	0.61	0.77	0.85	0.87	0.88	0.65	0.90	0.74	0.83	0.86	0.95	0.92	0.85	0.55	0.91	0.83
	mdok	0.70	0.85	0.99	0.82	0.63	0.50	0.59	0.94	0.90	0.47	0.96	0.61	0.88	0.80	0.97	0.97	0.85	0.47	0.94	0.81
hr	Qwen3-4B-Base	0.47	0.90	0.94	0.93	0.60	0.80	0.85	0.98	0.91	0.77	0.95	0.89	0.92	0.60	0.94	0.60	0.74	0.49	0.91	0.83
	XLM-R-large	0.72	0.84	0.87	0.83	0.78	0.66	0.74	0.96	0.84	0.74	0.86	0.67	0.92	0.79	0.91	0.47	0.75	0.65	0.85	0.79
	mdok	0.69	0.92	0.97	0.86	0.76	0.57	0.70	**0.99**	0.94	0.61	0.93	0.70	0.97	0.82	0.94	0.47	0.73	0.47	0.90	0.81
pl	Qwen3-4B-Base	0.47	0.62	0.97	0.89	0.60	0.78	**0.91**	0.96	0.83	0.77	0.97	**0.94**	0.97	0.51	0.97	0.98	0.71	0.47	0.94	0.83
	XLM-R-large	0.76	0.93	0.94	0.90	0.77	0.77	0.88	0.94	0.93	0.73	0.96	0.75	0.91	0.86	0.97	0.97	0.89	**0.83**	0.95	0.88
	mdok	0.78	**0.94**	0.98	0.81	**0.82**	0.58	0.84	0.97	**0.97**	0.79	**0.98**	0.85	0.97	**0.88**	0.97	0.98	**0.94**	0.47	0.96	**0.88**
sk	Qwen3-4B-Base	0.47	0.56	0.93	0.78	0.51	0.76	0.72	0.90	0.73	0.59	0.86	0.82	0.91	0.53	**1.00**	0.87	0.62	0.47	0.88	0.76
	XLM-R-large	0.61	0.62	0.84	0.74	0.54	0.50	0.53	0.55	0.53	0.48	0.61	0.69	0.57	0.72	0.99	0.56	0.55	0.51	0.72	0.64
	mdok	**0.81**	0.72	0.92	0.68	0.60	0.51	0.57	0.86	0.80	0.50	0.93	0.64	0.90	0.62	0.99	0.92	0.72	0.47	0.89	0.76
sl	Qwen3-4B-Base	0.47	0.87	0.96	0.89	0.67	**0.86**	0.87	0.97	0.93	**0.89**	0.94	0.92	0.96	0.53	0.97	0.99	0.75	0.47	0.95	0.86
	XLM-R-large	0.60	0.84	0.92	0.86	0.66	0.75	0.80	0.91	0.88	0.73	0.90	0.70	0.88	0.78	0.94	0.97	0.72	0.57	0.91	0.81
	mdok	0.75	0.93	0.97	0.93	0.79	0.67	0.82	0.96	0.94	0.85	0.90	0.84	0.95	0.84	0.97	**0.99**	0.89	0.47	0.95	0.88

Obr. 5 Porovnanie medzijazykovej prenositeľnosti binárnej klasifikácie (Macro F1) zvolených detekčných systémov pre jednotlivé testovacie jazyky v doméne novinových článkov.

Pri porovnaní s detektormi natrénovanými len na jednom jazyku môžeme pozorovať, že niektoré jednojazyčne dotrénované detektory (napr. špecializované na poľštinu) dosahujú vyššie macro F1 na kombinácii všetkých jazykov (All), a teda majú o niečo lepšiu prenositeľnosť do „neznámych" jazykov (teda nevidených počas dotrénovania, avšak v predtrénovaní prítomné).

Výsledky pre StatEnsembleLR na Obr. 6 potvrdzujú lepšie výsledky perlang alternatívy. Aj v tomto prípade ide o obidva sledované ukazovatele: vyššiu detekčnú schopnosť v jazykoch stredoeurópskeho regiónu, ako aj prenositeľnosť do ostatných jazykov. Alternatíva a), označená single, dosiahla najhoršiu detekčnú schopnosť v jazykoch stredoeurópskeho regiónu z porovnávaných architektúr, avšak pri prenositeľnosti do ostatných jazykov dopadla o niečo lepšie ako mean. Podobne ako v prípade dotrénovaných LLM modelov, aj v prípade StatEnsembleLR identifikujeme jednojazyčne natrénované detektory dosahujúce lepšie ukazovatele (najmä čeština a nemčina).

System		ar	bg	cs	de	el	en	es	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	All-CE	All
mean	StatEnsembleLR	0.47	0.53	0.56	0.60	0.45	0.48	0.54	0.58	0.58	0.58	0.64	0.56	0.74	0.53	0.63	0.66	0.55	0.56	0.61	0.57
perlang	StatEnsembleLR	0.47	0.53	0.65	**0.68**	0.45	0.48	0.54	0.70	0.70	0.58	0.59	0.56	0.74	0.53	**0.67**	0.66	0.55	0.56	0.67	0.59
single	StatEnsembleLR	0.57	0.56	0.53	0.56	0.51	0.47	0.51	0.56	0.59	0.54	0.62	0.52	0.69	0.59	0.67	0.61	0.57	0.61	0.60	0.58
cs	StatEnsembleLR	0.42	0.54	0.65	0.67	0.40	**0.58**	**0.63**	**0.73**	0.60	**0.68**	0.65	0.63	0.77	0.56	0.65	0.73	0.54	0.52	**0.67**	0.60
de	StatEnsembleLR	0.63	0.67	**0.67**	**0.68**	0.51	0.52	0.63	0.61	**0.71**	0.65	**0.72**	**0.66**	**0.79**	0.66	0.62	0.65	0.57	0.69	0.67	**0.65**
hr	StatEnsembleLR	0.39	0.49	0.57	0.63	0.37	0.55	0.59	0.70	0.56	0.64	0.63	0.60	0.68	0.51	0.60	**0.75**	0.46	0.52	0.64	0.57
hu	StatEnsembleLR	**0.66**	**0.71**	0.60	0.62	**0.63**	0.47	0.57	0.53	0.70	0.56	0.69	0.59	0.78	**0.71**	0.62	0.56	0.57	0.67	0.62	0.64
pl	StatEnsembleLR	0.60	0.60	0.53	0.56	0.48	0.47	0.52	0.51	0.58	0.51	0.59	0.53	0.73	0.59	0.57	0.59	**0.60**	**0.72**	0.56	0.58
sk	StatEnsembleLR	0.43	0.45	0.50	0.55	0.48	0.48	0.52	0.60	0.52	0.54	0.57	0.50	0.55	0.49	**0.67**	0.58	0.47	0.43	0.57	0.52
sl	StatEnsembleLR	0.39	0.44	0.49	0.53	0.41	0.47	0.51	0.54	0.47	0.54	0.52	0.51	0.52	0.45	0.54	0.66	0.44	0.44	0.54	0.49

Obr. 6 Porovnanie medzijazykovej prenositeľnosti binárnej klasifikácie (Macro F1) kombinácie štatistických detekčných systémov pre jednotlivé testovacie jazyky v doméne novinových článkov.

Pri osem-triednej klasifikácii (teda identifikácii konkrétneho generátora, ktorý je autorom textu) môžeme z výsledkov (Obr. 7) pozorovať výraznú prevahu perlang alternatívy architektúry detekčného systému (macro F1 0,89 v prípade mdok perlang vs. 0,83 v prípade mdok mean). Pri tomto porovnaní nedokázal žiadny z jednojazyčne dotrénovaných detektorov prekonať kombinovanú detekčnú schopnosť v jazykoch stredoeurópskeho regiónu, ako ani prenositeľnosť do ostatných jazykov. Ak si tieto výsledky porovnáme s multilingválne dotrénovanými detektormi (t. j. na všetkých 18 jazykoch datasetu kombinovane) v štúdii [2], ktorých detekčná schopnosť v jazykoch stredoeurópskeho regiónu dosiahla 0,94 macro F1 v prípade detektora mdok, tak môžeme predpokladať, že single riešenie by v tomto prípade dosiahlo lepšie výsledky. Kvôli inému počtu vzoriek v trénovaní však tieto dva systémy nie je možné objektívne porovnať a vyžadovalo by si to kontrolovaný experiment v budúcnosti.

System		ar	bg	cs	de	el	en	es	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	All-CE	All
	Qwen3-4B-Base	0.50	**0.73**	0.87	0.66	0.41	0.40	**0.58**	0.86	0.67	0.70	0.81	**0.70**	**0.82**	0.60	0.91	0.85	**0.66**	0.22	0.81	0.68
mean	XLM-R-large	0.43	0.66	0.72	0.46	0.43	0.23	0.35	0.70	0.67	0.44	0.66	0.37	0.52	0.52	0.80	0.75	0.58	0.30	0.69	0.55
	mdok	0.27	0.65	0.86	**0.74**	0.32	0.38	0.54	0.89	**0.76**	0.69	0.83	0.66	0.81	**0.63**	0.87	0.89	0.63	0.21	0.83	0.67
	Qwen3-4B-Base	0.50	**0.73**	0.92	0.66	0.41	0.40	**0.58**	0.94	0.67	0.70	0.89	**0.70**	**0.82**	0.60	**0.97**	0.95	**0.66**	0.22	0.86	**0.70**
perlang	XLM-R-large	0.43	0.66	0.76	0.46	0.43	0.23	0.35	0.76	0.67	0.44	0.77	0.37	0.52	0.52	0.89	0.76	0.58	0.30	0.73	0.57
	mdok	0.27	0.65	**0.94**	**0.74**	0.32	0.38	0.54	**0.95**	**0.76**	0.69	**0.92**	0.66	0.81	**0.63**	0.96	**0.95**	0.63	0.21	**0.89**	0.69
	Qwen3-4B-Base	0.48	0.66	0.92	0.65	0.37	0.37	0.52	0.80	0.67	0.61	0.78	0.63	0.79	0.58	0.90	0.83	0.64	0.18	0.80	0.66
cs	XLM-R-large	0.42	0.64	0.76	0.48	0.36	0.23	0.33	0.59	0.66	0.41	0.63	0.35	0.48	0.55	0.77	0.73	0.61	0.29	0.67	0.54
	mdok	0.27	0.49	**0.94**	0.69	0.24	0.32	0.47	0.81	0.74	0.61	0.83	0.58	0.74	0.59	0.85	0.83	0.62	0.23	0.81	0.63
	Qwen3-4B-Base	0.38	0.67	0.77	0.67	0.34	0.24	0.49	0.94	0.62	0.66	0.78	0.56	0.73	0.49	0.72	0.53	0.55	0.16	0.72	0.60
hr	XLM-R-large	0.34	0.56	0.63	0.39	0.36	0.13	0.23	0.76	0.58	0.38	0.57	0.28	0.50	0.44	0.51	0.15	0.50	0.28	0.54	0.45
	mdok	0.24	0.63	0.76	0.71	0.31	0.42	0.54	**0.95**	0.65	0.64	0.73	0.59	0.79	0.58	0.71	0.32	0.52	0.19	0.70	0.59
	Qwen3-4B-Base	0.43	0.57	0.73	0.64	0.38	**0.49**	0.56	0.72	0.55	0.67	0.89	0.63	0.68	0.50	0.71	0.69	0.57	0.18	0.72	0.61
pl	XLM-R-large	0.45	0.64	0.68	0.51	0.44	0.26	0.39	0.63	0.68	0.48	0.77	0.42	0.57	0.51	0.72	0.68	0.59	**0.41**	0.68	0.56
	mdok	0.27	0.61	0.80	0.71	0.35	0.39	0.55	0.80	0.71	**0.73**	**0.92**	0.64	0.80	0.61	0.74	0.80	0.64	0.21	0.78	0.65
	Qwen3-4B-Base	**0.58**	0.61	0.76	0.49	0.46	0.41	0.48	0.65	0.56	0.54	0.71	0.61	0.69	0.57	**0.97**	0.67	0.57	0.28	0.70	0.61
sk	XLM-R-large	0.48	0.49	0.52	0.40	**0.48**	0.19	0.27	0.40	0.41	0.29	0.46	0.39	0.33	0.49	0.89	0.45	0.43	0.24	0.53	0.45
	mdok	0.32	0.57	0.76	0.57	0.35	0.31	0.42	0.71	0.67	0.54	0.72	0.53	0.64	0.51	0.96	0.74	0.54	0.23	0.74	0.59
	Qwen3-4B-Base	0.36	0.71	0.83	0.54	0.38	0.35	0.52	0.80	0.70	0.60	0.72	0.64	0.70	0.54	0.82	0.95	0.60	0.16	0.77	0.63
sl	XLM-R-large	0.39	0.63	0.62	0.39	0.39	0.25	0.33	0.67	0.61	0.37	0.55	0.34	0.49	0.45	0.70	0.76	0.50	0.29	0.63	0.51
	mdok	0.21	0.57	0.76	0.72	0.27	0.43	0.51	0.76	0.62	0.68	0.69	0.55	0.68	0.51	0.72	**0.95**	0.53	0.18	0.75	0.60

Obr. 7 Porovnanie medzijazykovej prenositeľnosti viactriednej klasifikácie (Macro F1) zvolených detekčných systémov pre jednotlivé testovacie jazyky v doméne novinových článkov.

Naša hypotéza o lepšom single systéme však nebola potvrdená systémom StatEnsembleLR (Obr. 8), v ktorom taktiež dominuje perlang alternatíva ako v detekčnej schopnosti v jazykoch stredoeurópskeho regiónu, tak aj v prenositeľnosti do ostatných jazykov. Najhoršie výsledky dosiahla alternatíva mean, ktorá nedokázala prekonať ani niektoré jednojazyčne dotrénované detektory (napr. dotrénované v maďarčine).

System		ar	bg	cs	de	el	en	es	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	All-CE	All
mean	StatEnsembleLR	0.22	0.36	0.32	0.33	0.17	0.18	0.28	0.26	0.36	0.34	0.34	0.29	0.22	0.35	0.23	0.24	0.22	0.21	0.30	0.28
perlang	StatEnsembleLR	0.22	0.36	**0.35**	**0.40**	0.17	0.18	0.28	**0.31**	**0.46**	0.34	0.36	0.29	0.22	0.35	**0.39**	**0.30**	0.22	0.21	**0.38**	**0.32**
single	StatEnsembleLR	0.27	0.38	0.28	0.34	0.15	0.19	0.25	0.28	0.39	0.36	**0.38**	0.27	0.23	0.36	0.25	0.28	0.23	0.24	0.32	0.29
cs	StatEnsembleLR	0.16	0.33	**0.35**	0.27	0.16	0.18	0.29	0.26	0.31	0.30	0.28	0.28	**0.26**	0.31	0.22	0.19	0.23	0.19	0.27	0.26
de	StatEnsembleLR	**0.27**	0.39	0.28	**0.40**	0.19	**0.22**	**0.29**	0.29	0.39	**0.39**	0.37	**0.32**	0.23	**0.39**	0.22	0.20	0.28	**0.25**	0.31	0.31
hr	StatEnsembleLR	0.17	0.20	0.25	0.26	0.12	0.18	0.23	**0.31**	0.22	0.27	0.25	0.24	0.19	0.24	0.17	0.19	0.17	0.15	0.24	0.21
hu	StatEnsembleLR	0.24	**0.42**	0.25	0.30	**0.29**	0.17	0.26	0.19	**0.46**	0.30	0.36	0.26	0.21	0.39	0.24	0.24	**0.32**	0.25	0.30	0.30
pl	StatEnsembleLR	0.26	0.36	0.23	0.34	0.19	0.14	0.20	0.25	0.37	0.30	0.36	0.22	0.19	0.38	0.19	0.23	0.29	0.24	0.29	0.27
sk	StatEnsembleLR	0.15	0.21	0.21	0.19	0.23	0.04	0.12	0.16	0.23	0.19	0.20	0.14	0.13	0.15	**0.39**	0.18	0.13	0.12	0.23	0.19
sl	StatEnsembleLR	0.12	0.20	0.19	0.23	0.11	0.13	0.16	0.17	0.22	0.21	0.24	0.19	0.15	0.20	0.19	**0.30**	0.17	0.16	0.22	0.19

Obr. 8 Porovnanie medzijazykovej prenositeľnosti viactriednej klasifikácie (Macro F1) kombinácie štatistických detekčných systémov pre jednotlivé testovacie jazyky v doméne novinových článkov.

Z dosiahnutých výsledkov jednoznačne vyplýva, že v doméne novinových článkov a viactriednej klasifikácii je alternatíva perlang najvhodnejšia z hľadiska maximalizácie detekčnej schopnosti v jazykoch stredoeurópskeho regiónu, tak aj v prenositeľnosti do ostatných jazykov. Pri porovnaní s výsledkami z Kap. 2.1 vidíme rozdiel, ktorý je spôsobený

inou kombináciou jednojazyčne špecializovaných detektorov do viacjazyčného riešenia. V predchádzajúcej časti bola kombinácia realizovaná pomocou priemerovania pravdepodobností, pričom v tejto časti sme použili väčšinovú voľbu. Takže perlang riešenie s väčšinovou voľbou je najvhodnejšou alternatívou len v prípade nedostupnosti hodnôt pravdepodobností pre jednotlivé triedy.

2.3 Výpočtová náročnosť

Pri nasadení detekčného systému v reálnych podmienkach je výpočtová náročnosť kľúčovým faktorom, ktorý ovplyvňuje praktickú využiteľnosť jednotlivých architektúr. V tejto časi porovnávame tri sledované architektúry (single, mean, perlang) z hľadiska pamäťových nárokov, doby inferencie a škálovateľnosti systému.

Pamäťové nároky. Architektúra single vyžaduje dotrénovanie a uloženie len jedného modelu, čo predstavuje optimálne pamäťové nároky (napr. pre mDeBERTa-v3-base ide o cca 300 miliónov parametrov, teda cca 550 MB). Naopak, architektúry mean a perlang vyžadujú udržiavanie všetkých piatich (resp. siedmich) jednojazyčne dotrénovaných modelov, čo v prípade paralelného načítania zvyšuje pamäťové nároky lineárne s počtom jazykov (5-7×). Pri sekvenčnom načítaní modelov je pamäťový odtlačok porovnateľný s alternatívou single, avšak za cenu výrazne vyššej doby inferencie (oneskorenie výsledkov).

Doba inferencie. Pri architektúre single prebehne pre každý vstupný text práve jedna inferencia, čo predstavuje najnižšiu latenciu. Architektúra perlang taktiež vykoná pre texty v známych jazykoch jednu inferenciu (konkrétneho modelu špecializovaného na daný jazyk), výsledkom čoho je latencia porovnateľná s architektúrou single. Architektúra mean je v tomto ohľade najnáročnejšia, keďže pre každý text vyžaduje inferenciu cez všetkých päť (resp. sedem) modelov, čiže celkový čas spracovania narastá minimálne päťnásobne oproti single architektúre.

Škálovateľnosť. Pridanie podpory nového jazyka si pri architektúre single vyžaduje opätovné dotrénovanie celého modelu na rozšírenej kombinácii jazykov, čo je výpočtovo nákladné. Prípadne je možné sa spoľahnúť na lepšiu jazykovú prenositeľnosť do iných jazykov (kvôli viacjazyčnému dotrénovaniu pôvodného modelu). Oproti tomu architektúry mean a perlang umožňujú pridanie nového jednojazyčného detektora bez zásahu do existujúcich modelov, čo je z hľadiska modulárnosti a inkrementálneho rozširovania výhodnejšie.

Z tejto analýzy teda vyplýva, že z výpočtového hľadiska predstavuje najúspornejšiu voľbu architektúra single (úložisko, pamäť, latencia). Pre moderné systémy s dostupnou viacjadrovou (resp. multi-GPU) infraštruktúrou a požiadavkou na nízku latenciu je architektúra perlang výhodnou voľbou, keďže kombinuje nízku dobu inferencie pri známych jazykoch s modularitou (ľahká rozšíriteľnosť pridaním ďalšieho špecializovaného modelu pre nový jazyk). Architektúra mean je z výpočtového hľadiska najnákladnejšia a nie je odporúčaná pre systémy s obmedzenou výpočtovou kapacitou alebo prísnou požiadavkou na nízku latenciu.

3 Záver

Táto výskumná správa reprezentuje porovnávaciu štúdiu troch architektúr systémov pre detekciu MGT v stredoeurópskom multilingválnom prostredí: a) jedného viacjazyčne dotrénovaného modelu (single), b) kombinácie jednojazyčne dotrénovaných modelov pomocou priemerovania pravdepodobností predikcie alebo väčšinovej voľby (mean) a c) kombinácie jednojazyčne dotrénovaných modelov pomocou preferencie jazykovo korešpondujúceho detektora (perlang). Z hľadiska detekčnej úspešnosti pri binárnej klasifikácii na datasete CEAID (dve textové domény) dosiahla pri troch zo štyroch testovaných základných modelov najlepšie výsledky architektúra mean (priemerovanie pravdepodobností), čo vedie k jej odporúčaniu v situáciách, keď sú hodnoty pravdepodobností pre jednotlivé triedy dostupné. V kontexte robustnosti pri viactriednej klasifikácii a medzijazykovej prenositeľnosti, testovanej na datasete MULTITuDE_v3 naprieč 18 jazykmi, jednoznačne dominovala architektúra perlang ako v jazykoch stredoeurópskeho regiónu, tak aj pri prenositeľnosti do neznámych jazykov (mimo dotrénovacích). Z pohľadu výpočtovej náročnosti je architektúra single najúspornejšia z hľadiska pamäte a doby inferencie. Architektúra perlang predstavuje vyvážený kompromis medzi detekčnou výkonnosťou, modularitou a latenciou. Architektúra mean je najnákladnejšia a pre systémy s obmedzenými zdrojmi sa neodporúča.

Na základe súhrnu všetkých sledovaných dimenzií hodnotíme architektúru perlang ako najvhodnejšiu pre nasadenie MGT detekčného systému v stredoeurópskom multilingválnom prostredí. Jej modulárna povaha umožňuje flexibilné rozširovanie o nové jazyky bez nutnosti opätovného trénovania celého systému, pričom v prípadoch, keď sú k dispozícii hodnoty pravdepodobností, možno zvýšiť výkonnosť nahradením väčšinovej voľby priemerovaním pravdepodobností (overené len pri binárnej klasifikácii). Budúci výskum by sa mal zamerať na kontrolovaný experiment porovnávajúci architektúru perlang s viacjazyčne dotrénovaným modelom single pri rovnakom počte trénovacích vzoriek, a to predovšetkým v scenári viactriednej klasifikácie. Ďalším smerom výskumu je skúmanie efektívnejších metód fúzie jednojazyčných detektorov a nasadenie multilingválneho detektora pre neznáme jazyky (potenciál lepšej prenositeľnosti).

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. Accepted to Findings of ACL 2026.

[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. Accepted to ACL 2026.