V3.3 Výskumná správa o obohatených datasetoch pre trénovanie modelov v oblasti výskumu kvality informácií

Názov projektu	**Generovanie personalizovaného obsahu vo výskume kvality informácií**
Akronym	**GEPERO**
Kód projektu	**09I01-03-V04-00068**
Začiatok projektu	**01. 11. 2024**
Trvanie projektu	**20 mesiacov**

Financované Európskou úniou NextGenerationEU

PLÁN OBNOVY

Obsah

1 Úvod

Obohatené datasety (angl. data augmentation) pre trénovanie jazykových modelov predstavujú kľúčový prvok pri zlepšovaní ich schopností v oblasti hodnotenia kvality informácií. Jednou z dôležitých oblastí skúmania kvality textov je úloha detekcie strojovo-generovaného textu (MGT, angl. machine-generated text), najmä kvôli kontinuálne sa zlepšujúcej schopnosti veľkých jazykových modelov (LLM, angl. large language model) rýchlo a jednoducho generovať text, ktorý je pre človeka ťažko rozoznateľný od autentického textu písaného človekom.

Táto výskumná správa nadväzuje na predchádzajúce výskumné správy (V3.1 a V3.2) a opisuje tvorbu, aplikáciu a vyhodnocovanie obohatených datasetov o vygenerované personalizované texty, vytvorených v rámci projektu GEPERO. Tieto datasety vychádzajú zo štúdií a poznatkov získaných v predchádzajúcich fázach výskumu a sú aplikované na vzorovú úlohu detekcie MGT, ktorá predstavuje pomocný indikátor použiteľný napr. pri detekcii dezinformácií. V štúdiách [1] a [3] sme identifikovali štatisticky významný vplyv použitia personalizácie pri generovaní textu na jeho detegovateľnosť. V tejto výskumnej správe opisujeme našu štúdiu vplyvu obohatenia tréningového datasetu o personalizovane generované texty na jeho schopnosti detekcie personalizovane generovaných textov v doméne novinových článkov ako aj textov sociálnych sietí.

Táto výskumná správa je rozdelená na tri časti. V prvej časti (Kap. 2.1) opisujeme metodológiu tvorby obohatených datasetov. V druhej časti (Kap. 2.2) sa venujeme overeniu vplyvu ich použitia na všeobecnú detekčnú schopnosť, či už v rámci domény, alebo mimo trénovacej domény. Tretia časť (Kap. 2.3) sa venuje vyhodnoteniu vplyvu ich použitia na detegovateľnosť personalizovaných dezinformačných textov.

2 Obohatenie datasetov pomocou generovaného personalizovaného textu

Keďže náš predchádzajúci výskum [1, 3] potvrdil štatisticky významný vplyv použitia personalizácie pri generovaní textu na jeho detegovateľnosť, cieľom tejto časti výskumu je preskúmať vplyvu obohatenia tréningového datasetu MGT detektorov o personalizovane generované texty na jeho schopnosti detekcie personalizovane generovaných textov.

2.1 Obohacujúce datasety personalizovaných MGT

Ako datasety obohacujúce trénovaciu množinu textov pre trénovanie MGT detektorov použijeme datasety a ich podmnožiny vytvorené v rámci projektu GEPERO, konkrétne dataset PerQ [2] (použitý na vyhodnotenie kvality prispôsobenia textu pre platformu sociálnych sietí) a dataset PerComm (predstavený v predchádzajúcej výskumnej správe V3.2 na vyhodnotenie vplyvu parameterov generovania textu na kvalitu personalizácie).

2.1.1 PerQ

PerQ dataset [2] sme vytvorili v rámci projektu GEPERO na vyhodnotenie schopnosti jazykových modelov personalizovať texty vo viacerých jazykoch pre zvolené cieľové platformy sociálnych sietí. Dataset je v súčasnosti zverejnený na platforme Zenodo[1] pre nekomerčné výskumné účely. Ako zdrojové texty sme použili existujúce novinové články obsiahnuté v datasete MassiveSumm, pričom titulok článku bol použitý v inštrukcii na generovanie textu a telo článku v inštrukcii na modifikáciu textu. Takto bola zabezpečená konzistencia v témach týchto dvoch skupín textov (generované vs. modifikované). Výsledný dataset obsahuje 7 jazykov, zahŕňajúcich angličtinu, francúzštinu, maďarčinu, nemčinu, ruštinu, slovenčinu a taliančinu. Ako cieľové platformy boli zvolené Twitter/X, Telegram a Signal. Do porovnania sme zámerne zahrnuli 2 veľkosti modelov 3 rovnakých architektúr, konkrétne Gemma-3-27B-IT, Gemma-3-4B-IT, Llama-3.3-70B-Instruct, Llama-3.2-3B-Instruct, Qwen3-32B a Qwen3-1.7B.

Kvalita personalizácie bola vyhodnotená pomocou 3 LLM (Mistral-Small-3.1-24B, Aya-Expanse-32B a QwQ-32B) a väčšinová voľba zabepečila identifikáciu generátorov s najlepšou

[1] https://doi.org/10.5281/zenodo.20919707

mierou personalizácie voči cieľovej platforme, ako aj umožnila filtráciu textov s lepšou kvalitou personalizácie. Týmto spôsobom sme vytvorili 3 obohacujúce podmnožiny datasetu PerQ:

- PerQ – všetky vygenerované texty, zahŕňajúce aj dobrú aj zlú kvalitu personalizácie textov (výhodou je vyššia diverzita generovaných textov).
- PerQ-bestquality – zo 4-stupňovej škály hodnotenia kvality personalizácie boli vybrané texty dosahujúce aspoň druhú najvyššiu kvalitu podľa väčšinovej voľby.
- PerQ-bestpersonalizer – model Gemma-3-4B-IT poskytuje dobrý kompromis medzi výslednou kvalitou personalizácie vygenerovaných textov a výpočtovou efektívnosťou v prípade použitia ako samostatného dodatočného obohacujúceho modelu (angl. augmentator).

2.1.2 PerComm

PerComm (V3.2) dataset sme vytvorili v rámci projektu GEPERO na vyhodnotenie vplyvu parametrov generovania textu na výslednú kvalitu personalizácie vygenerovaného textu. Ako zdrojové texty sme použili vzorky textov dostupné v datasete Common Corpus (Langlais et al., 2025), ktorý zabezpečuje vysokú diverzitu jazykov aj typov textov. Každý z vybraných textov bol použitý v tvorbe práve jednej požiadavky (vstupnej inštrukcie), pre ktorú bola pseudonáhodne zvolená práve jedna kombinácia týchto parametrov z daných množín hodnôt:

- Akcia vykonania prispôsobenia vstupného textu: modifikácia, sumarizácia, prepis, parafrázovanie, pokračovanie
- Parameter temperature: 0,0; 0,25; 0,5; 0,75; 1,0
- Parameter top_p: 0,1; 0,25; 0,5; 0,75; 0,95; 1,0
- Parameter top_k: 5, 30, 50, 70, 100
- Parameter repetition_penalty: 0,8; 0,9; 1,0; 1,1; 1,2

Celkovo bolo teda použitých 3 750 možných kombinácií parametrov generovania textu. V každej požiadavke bol tento vstupný text personalizovaný pre jednu z nasledujúcich možnosti (taktiež zvolenú pseudonáhodne): žiadna (None, ako kontrolná vzorka nepersonalizovaných textov), platforma Telegram, platforma Twitter, platforma Mastodon, študenti, rodičia, seniori, konzervatívci, liberáli. Tieto ciele personalizácie boli špecifikované len identifikátorom. Tieto vstupné inštrukcie boli použité každým zo 6 zvolených LLM: GPT-oss-20B, Gemma-3-27B-it, Qwen3-32B, Mistral-Small-3.1-24B-Instruct-2503, Phi-4 (14B) a DeepSeek-R1-Distill-Qwen-32B. Takto vytvorený dataset s veľkou diverzitou jazykov (25 písaných ľudských jazykov a v menšom zastúpení počtu vzoriek aj 23 programovacích

jazykov textov zdrojových kódov), domén (vládne zdroje, veda, Wikipedia, kód, atď.), parametrov generovania (3 750 možných kombinácií), ako aj diverzitou generátorov (6 LLM rôznych architektúr a veľkostí) môže byť sám o sebe použitý na trénovanie detektorov (keďže obsahuje aj veľké množstvo pôvodných ľudských textov), ale taktiež ako doplnkový obohacujúci dataset obsahujúci personalizované aj nepersonalizované texty.

V Tab. 1 je zobrazený prehľad vyššie opísaných datasetov určených na obohatenie trénovacích datasetov pre detekciu MGT, pričom zahŕňame do porovnania aj kombináciu PerQ a PerComm datasetov.

Dataset	LLM	Jazyk	MGT	Ľudské texty
PerQ	6	7	25 181	0
PerQ-bestquality	6	7	15 495	0
PerQ-bestpersonalizer	1	7	4 196	0
PerComm	7	25 ľudských + 23 programových	109 992	41 637
PerQ + PerComm	11	25 ľudských + 23 programových	135 173	41 637

Tab. 1 Štatistický prehľad diverzity obohacujúcich datasetov.

2.2 Vplyv na detekčnú schopnosť MGT detektorov

Na overenie vplyvu obohatenia tréningovej množiny MGT detektorov o generované personalizované texty na ich výslednú detekčnú schopnosť použijeme ako referenčný model (baseline) robustne dotrénovaný viacjazyčný model založený na mDeBERTa-v3-base (Macko et al., 2025). Kvôli sledovaniu vplyvu na medzi-doménovú transferabilitu detekčných schopností sme obmedzili trénovaciu sadu na doménu novinových článkov. Trénovacia sada teda obsahuje 136 577 ľudských textov a 143 527 MGT textov z trénovacej časti datasetu MULTITuDE_v3 (7 LLM, 22 jazykov) a 120 090 obfuskovaných (t. j. adverzariálne upravených s účelom zabránenia detekcie) MGT textov z datasetu MULTITuDE_v2 (8 LLM, 3 jazyky). Celkovo trénovací dataset pokrýva doménu novinových článkov (oba vyššie uvedené datasety vychádzajú z titulkov novinových článkov v datasete MassiveSumm, podobne ako PerQ), 13 LLM (Alpaca-LoRA-30B, Aya-101, GPT3-text-davinci-003, GPT-3.5-Turbo, GPT-4, Llama-65B , Llama-2-70B-chat-hf, Mistral-7B-Instruct-v0.2, OPT-66B, OPT-IML-Max-1.3B, OPT-IML-Max-30B, v5-Eagle-7B-HF, Vicuna-13B) a 22 jazykov.

Takto natrénovaný detektor je následne vyhodnotený na testovacích častiach existujúcich datasetov MULTITuDE_v3 (novinové články) a MultiSocial (texty z 5 rôznych platforiem sociálnych sietí), pričom oba majú rovnaké zastúpenie 22 jazykov (ako v trénovacej množine). Pomocou obohatenia (pridania dátových vzoriek) a zamiešania tréningovej sady o jednotlivé

datasety z Tab. 1 sme dotrénovali ďalších 5 verzií modelov a vyhodnotili na rovnakých testovacích datasetoch. Na vyhodnotenie sme zvolili metriku AUC ROC, ktorá reprezentuje všeobecnú detekčnú schopnosť (bez stanovenia konkrétneho prahu pre klasifikáciu medzi dvoma triedami). Keďže trénovacia množina referenčného modelu obsahuje približne 400 tis. vzoriek, obmedzili sme trénovanie aj obohatených verzii modelov na 400 tis. videných vzoriek počas tréningu. Týmto sme efektívne zamedzili vplyvu väčšej trénovacej množiny na vyhodnotenie.

Výsledky v Tab. 2 indikujú, že pre všeobecnú detekciu LLM generovaného textu je dôležitejšia diverzita generátorov a štýlov textov (zabezpečená obohatením o dataset PerComm) ako samotná prítomnosť personalizácie. Pri detailnejšom pohľade na jednotlivé domény reprezentované týmito dvomi datasetmi sa ukázali výrazné rozdiely. V doméne novinových článkov dosahovali všetky modely veľmi vysoké výsledky s AUC približne 0,99. Rozdiely medzi jednotlivými variantmi boli minimálne, čo naznačuje, že základná kombinácia tréningových dát referenčného modelu poskytuje dostatočné pokrytie tejto domény a obohatenie o personalizované dáta neprináša výrazné zvýšenie detekčnej schopnosti.

Detektor	MULTITuDE_v3	MultiSocial
baseline	0,9872	0,7795
baseline + PerQ	0,9882	0,7753
baseline + PerQ-bestquality	**0,9886**	0,7821
baseline + PerQ-bestpersonalizer	0,9879	0,7835
baseline + PerComm	0,9870	**0,8154**
baseline + PerQ + PerComm	0,9875	0,8132

Tab. 2 Všeobecná detekčná schopnosť (AUC ROC) zvolených MGT detektorov na testovacích častiach evaluačných datasetov.

Naopak pri sociálnych sieťach boli rozdiely výraznejšie. Zatiaľ čo PerQ dáta nepriniesli konzistentné zlepšenie (hoci cieľom personalizácie boli výlučne platformy sociálnych sietí), PerComm dataset zvýšil AUC ROC približne o 3 až 4 %. Tento výsledok poukazuje na vyššiu medzi-doménovú transferabilitu pri použití rozmanitejších generátorov a rôznych typov textových transformácií. DeLong test ($p < 0,05$) potvrdil štatistickú významnosť väčšiny rozdielov, dokonca aj pridanie PerQ datasetu vyhodnotené na MULTITuDE_v3 v porovnaní s referenčným detektorom.

Výsledky v Tab. 3 ukazujú, že obohacujúci dataset PerComm zvýšil robustnosť detektora konzistentne vo všetkých testovaných jazykoch (pri zmene textovej domény). Hoci všetky testované jazyky boli zahrnuté do tréningu aj referenčného detektora, zmena formátu a štýlu

textov pri prechode do domény sociálnych sietí rapídne ovplyvnila jeho detekčné schopnosti. Tieto sú síce znížene aj v prípade dotrénovania na PerComm dátach, ich štýlova diverzita zabezpečila, že pokles úspešnosti detekcie je o niečo nižší.

Detektor	ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	ALL
baseline	0,87	0,84	0,77	0,72	0,81	0,62	0,83	0,80	0,78	0,79	0,80	0,70	0,67	0,83	0,72	0,77	0,73	0,81	0,75	0,84	0,82	0,87	0,78
baseline + PerQ	0,85	0,82	0,74	0,76	0,79	0,62	0,80	0,79	0,80	0,79	0,80	0,73	0,72	0,81	0,74	0,77	0,74	0,79	0,75	0,84	0,83	0,84	0,78
baseline + PerQ-bestquality	0,85	0,83	0,75	0,77	0,79	0,64	0,81	0,80	0,82	0,80	0,82	0,73	0,73	0,82	0,74	0,78	0,74	0,79	0,77	0,84	0,83	0,85	0,78
baseline + PerQ-bestpersonalizer	0,87	0,84	0,78	0,74	0,81	0,64	0,82	0,80	0,80	0,80	0,80	0,72	0,70	0,83	0,72	0,77	0,74	0,81	0,76	0,85	0,82	0,86	0,78
baseline + PerComm	**0,89**	**0,86**	**0,80**	0,79	**0,83**	**0,73**	**0,85**	**0,84**	0,82	**0,82**	**0,84**	0,76	0,74	**0,84**	0,78	0,80	**0,78**	**0,83**	**0,81**	**0,87**	0,84	**0,88**	**0,82**
baseline + PerQ + PerComm	0,87	0,86	0,78	**0,80**	0,83	0,73	0,84	0,84	**0,82**	0,81	0,84	**0,77**	**0,76**	0,84	**0,79**	**0,80**	0,77	0,82	0,80	0,87	**0,85**	0,87	0,81

Tab. 3 Všeobecná detekčná schopnosť (AUC ROC) zvolených MGT detektorov v jednotlivých jazykoch v doméne sociálnych sietí.

2.3 Vplyv na detegovateľnosť personalizovaných dezinformačných textov generovaných pomocou LLM

Na overenie vplyvu obohatenia tréningovej množiny MGT detektorov o generované personalizované texty na detegovateľnosť vygenerovaných personalizovaných dezinformačných textov použijeme všetkých 6 verzií detektorov dotrénovaných ako bolo uvedené v Kap. 2.2 (t. j. bez obohatenia tréningovej množiny a s obohatením tréningovej množiny rôznymi obohacujúcimi datasetmi podľa Tab. 1). Následne pomocou týchto detektorov vyhodnotíme mieru detekcie vygenerovaných textov obsiahnutých v našich datasetoch PerDisNews [1] a mPerDisSocial [3] (obsahujúcich personalizované aj nepersonalizované texty vo forme novinových článkov aj textov sociálnych sietí) pomocou metriky TPR (angl. true positive rate), t. j. koľko percent zo všetkých vygenerovaných textov bolo správne identifikovaných ako MGT.

Dataset **PerDisNews** [1] sme vytvorili v rámci projektu GEPERO na vyhodnotenie základnej miery personalizačných schopností LLM modelov (anglické novinové články). Dataset je v súčasnosti zverejnený na platforme Zenodo[2] pre nekomerčné výskumné účely. V datasete je zahrnutá personalizácia pre 7 cieľových skupín z 3 kategórií: politická afiliácia (európski konzervatívci a európski liberáli), bývanie (mesto, dedina), vek (študenti, rodičia, seniori). Dataset porkýva 6 rôznych dezinformačných naratívov z oblasti zdravia a politiky. Texty sú vygenerované pomocou 6 rôznych LLM modelov (rôzne architektúry a veľkosti): Falcon-40B, Vicuna-33B, GPT-4o, Gemma-2-27B, Llama-3.1-70B a Mistral-Nemo. Použité boli 3 typy

[2] https://doi.org/10.5281/zenodo.15463489

inštrukcií: bez špecifikácie cieľovej skupiny (t. j. bez požiadavky na personalizáciu - kvôli porovnaniu), s jednoduchou špecifikáciou cieľovej skupiny (len názov skupiny) a s detailnou špecifikáciou cieľovej skupiny (poskytnutím stručnej charakteristiky). Finálny dataset takto vygenerovaných textov je nazvaný PerDisNews a obsahuje 2238 dezinformačných textov.

Dataset **mPerDisSocial** [3] sme vytvorili v rámci projektu GEPERO na vyhodnotenie všeobecných personalizačných schopností viacjazyčných LLM modelov. Aj tento dataset je v súčasnosti zverejnený na platforme Zenodo[3] pre nekomerčné výskumné účely. Vychádzali sme z metodológie tvorby datasetu PerDisNews, ktorú sme však rozšírili vo viacerých aspektoch. V datasete je síce použitá rovnaká množina 6 dezinformačných naratívov z dvoch oblastí (zdravie a politika), ale kvôli redukcii dimenzionality sme obmedzili počet cieľových skupín na dve ("európski konzervatívci" s najvyššou a "mestskí obyvatelia" s najnižšou nameranou kvalitou personalizácie) pri použití detailnej špecifikácie cieľovej skupiny. Pre účely porovnania sú v datasete zahrnuté aj vzorky bez požiadavky na personalizáciu pre cieľovú skupinu. Tieto vzorky boli pri generovaní požadované personalizovať pre 3 cieľové platformy (Mastodon, Telegram a Twitter/X). Dataset pokrýva 10 jazykov s dôrazom na jazyky regiónu centrálnej Európy (čeština, chorvátčina, maďarčina, nemčina, poľština, slovenčina a slovínčina), okrem ktorých obsahuje aj angličtinu, estónčinu a ukrajinčinu. Kvôli porovnaniu pozitívneho a negatívneho využitia personalizačných schopností jazykových modelov sú v datasete pre každý každý dezinformačný naratív vygenerované vzorky proti naratívu a podporujúce naratív. Celkovo je v datasete pokrytých 6 x 3 x 3 x 10 x 2 = 1080 kombinácií čŕt v požiadavkách na generovanie textov (17 278). Texty sú rovnomerne vygenerované 16 verziami novších aj starších LLM rôznych architektúr a veľkostí (DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-32B, Gemma-2-27B-it, Gemma-2-2B-it, Gemma-2-9B-it, Gemma-3-27B-it, Gemma-3-4B-it, Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct, Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Llama-3.3-70B-Instruct, Mistral-Nemo-Instruct-2407, Qwen3-1.7B, Qwen3-32B, Qwen3-4B).

Výsledky na datasete PerDisNews (Tab. 4) ukazujú, že personalizované novinové články sú pre detektory veľmi dobre rozpoznateľné. Už referenčný model identifikuje viac ako 96 % vzoriek. Všetky augmentované detektory prekračujú 99 % úspešnosť detekcie. Zároveň výsledky potvrdzujú, že pre referenčný detektor je vyšší stupeň personalizácie náročnejší na detekciu (<97 % pri detailnej špecifikácii cieľovej skupiny vs. >99 % bez personalizácie). Táto zraniteľnosť detektora je úspešne odstránená po obohatení tréningovej množiny detektora

[3] https://doi.org/10.5281/zenodo.20398148

ľubovoľným so zvolených obohacujúcich datasetov (dokonca aj pridanie 4 tis. textov jedného personalizátora postačuje).

Detektor	Žiadna	Jednoduchá	Detailná
baseline	0,9960	0,9749	0,9696
baseline + PerQ	**1,0000**	0,9987	**1,0000**
baseline + PerQ-bestquality	**1,0000**	**1,0000**	**1,0000**
baseline + PerQ-bestpersonalizer	**1,0000**	0,9907	0,9974
baseline + PerComm	**1,0000**	**1,0000**	**1,0000**
baseline + PerQ + PerComm	**1,0000**	**1,0000**	**1,0000**

Tab. 4 Úspešnosť detekcie (TPR) vygenerovaných textov zvolených MGT detektorov v datasete PerDisNews podľa špecifickosti opisu cieľa personalizácie.

Keďže pri takejto vysokej úspešnosti detekcie všetkých detektorov sú rozdiely veľmi malé, analyzovali sme ich schopnosť detekcie pri zohľadnení pravdepodobnosti s akou predikujú danú vzorku za MGT. Pri tomto vyhodnotení teda považujeme text sa úspešne detegovaný len ak detektor predikoval MGT s pravdepodobnosťou 100 % (teda plná konfidenčná hodnota). Výsledky takéhoto porovnania sú uvedené v Tab. 5, kde sú rozdiely už jednoznačne viditeľné. Najväčší prínos k „istote“ detektora má augmentácia pomocou PerComm datasetu, najmenší (ale stále veľmi výrazný) má pridanie celého PerQ datasetu do trénovacej množiny. Vo všetkých prípadoch je stále viditeľný vplyv personalizácie na mieru „istoty“ MGT detektora, keďže detailná špecifikácia cieľovej skupiny pri každom detektore konzistentne zapríčinila najnižšiu úspešnosť takejto detekcie.

Detektor	Žiadna	Jednoduchá	Detailná
baseline	0,7553	0,7249	0,6720
baseline + PerQ	0,8783	0,8386	0,8135
baseline + PerQ-bestquality	0,9074	0,8664	0,8492
baseline + PerQ-bestpersonalizer	0,8915	0,8426	0,8267
baseline + PerComm	**0,9193**	**0,9048**	**0,8995**
baseline + PerQ + PerComm	0,8862	0,8862	0,8677

Tab. 5 Úspešnosť detekcie (TPR) vygenerovaných textov zvolených MGT detektorov pri použití predikcie s plnou konfidenčnou hodnotou v datasete PerDisNews podľa špecifickosti opisu cieľa personalizácie.

Výsledky na datasete mPerDisSocial podľa cieľových skupín ukazujú, či už pri predvolenej predikcii (Tab. 6), ako aj pri predikcii s plnou konfidenčnou hodnotou (Tab. 7), že augmentácia pomocou celého PerQ datasetu pomohla úspešnosti predikcie najviac. Z tohto výsledku jednoznačne vyplýva, že doplnenie očakávanej formy textu do trénovania (aj keď len vo forme generovaných dát) pomáha následnej úspešnosti detekcie MGT v danej doméne. Keďže boli

pridané len generované texty prispôsobené do formy a štýlu sociálnych sietí, ktoré doplnili novinové články, existuje obava, že každý text vo forme sociálnych sietí by mohol byť predikovaný ako MGT. Toto je však vylúčené na základe výsledkov hodnôt AUC ROC týchto detektorov na datasete MultiSocial (Tab. 2 a Tab. 3), ktorý obsahuje len texty sociálnych sietí. Keďže sú tieto hodnoty vysoko nad 0,5, môžeme tento negatívny jav definitívne vylúčiť.

Detektor	Žiadna	Konzervatívci	Mešťania
baseline	0,9316	0,9410	0,9200
baseline + PerQ	**0,9990**	0,9993	**0,9993**
baseline + PerQ-bestquality	0,9920	0,9927	0,9889
baseline + PerQ-bestpersonalizer	0,9663	0,9693	0,9653
baseline + PerComm	0,9821	0,9839	0,9812
baseline + PerQ + PerComm	0,9988	**0,9995**	0,9990

Tab. 6 Úspešnosť detekcie (TPR) vygenerovaných textov zvolených MGT detektorov v datasete mPerDisSocial podľa cieľovej skupiny personalizácie.

Detektor	Žiadna	Konzervatívci	Mešťania
baseline	0,3318	0,3564	0,3295
baseline + PerQ	**0,6078**	**0,6373**	**0,6418**
baseline + PerQ-bestquality	0,5385	0,5477	0,5488
baseline + PerQ-bestpersonalizer	0,4740	0,4771	0,4698
baseline + PerComm	0,4719	0,5097	0,4988
baseline + PerQ + PerComm	0,5137	0,5566	0,5682

Tab. 7 Úspešnosť detekcie (TPR) vygenerovaných textov zvolených MGT detektorov pri použití predikcie s plnou konfidenčnou hodnotou v datasete mPerDisSocial podľa cieľovej skupiny personalizácie.

Tab. 8 a Tab. 9 podobne zobrazujú výsledky porovnania na datasete mPerDisSocial podľa cieľových platforiem sociálnych sietí. Opäť môžeme pozorovať jednoznačnú dominanciu detektorov dotrénovaných celom PerQ datasete. Použitie len dát z najlepšieho personalizátora je v tomto prípade nie optimálne (hoci stále výrazne zvyšuje úspešnosť detekcie oproti referenčnému detektoru), vzhľadom na nízky počet týchto vzoriek (4 tis.) v provnaní s pôvodným trénovacím datasetom (400 tis.).

Detektor	Mastodon	Telegram	Twitter
baseline	0,9295	0,9405	0,9226
baseline + PerQ	**0,9991**	**0,9993**	**0,9991**
baseline + PerQ-bestquality	0,9899	0,9925	0,9911
baseline + PerQ-bestpersonalizer	0,9637	0,9743	0,9628
baseline + PerComm	0,9826	0,9832	0,9814
baseline + PerQ + PerComm	**0,9991**	0,9991	0,9990

Tab. 8 Úspešnosť detekcie (TPR) vygenerovaných textov zvolených MGT detektorov v datasete mPerDisSocial podľa cieľovej platformy personalizácie.

Detektor	Mastodon	Telegram	Twitter
baseline	0,3431	0,4151	0,2595
baseline + PerQ	**0,6352**	**0,7033**	**0,5484**
baseline + PerQ-bestquality	0,5500	0,6099	0,4752
baseline + PerQ-bestpersonalizer	0,4724	0,5517	0,3967
baseline + PerComm	0,4998	0,5568	0,4238
baseline + PerQ + PerComm	0,5431	0,6165	0,4790

Tab. 9 Úspešnosť detekcie (TPR) vygenerovaných textov zvolených MGT detektorov pri použití predikcie s plnou konfidenčnou hodnotou v datasete mPerDisSocial podľa cieľovej platformy personalizácie.

Výsledky tohto výskumu teda ukázali, že prítomnosť aj malého počtu personalizovaných textov v trénovacej množine môže zásadne ovplyvniť schopnosť detekcie personalizovaných dezinformačných textov. Avšak vyhodnotenie úspešnosti MGT detekcie na textovej doméne odlišnej od majoritnej trénovacej domény ukázal, že prítomnosť textov cieľovej domény (sociálnych sietí v našom prípade) v trénovacej množine je pri detekcii MGT v tejto doméne kľúčová, pričom samotná diverzita generátorov a štýlov textov (dodaných datasetom PerComm) nestačí ako náhrada za doménovo-relevantné vzorky.

3 Záver

Náš výskum v oblasti obohacovania datasetov pre trénovanie MGT detektorov o personalizované vzorky a vplyv takéhoto obohacovania na ich detekčnú schopnosť priniesol nové zaujímavé výsledky pre túto úlohu. Výsledky celkovo potvrdzujú, že obohatenie trénovacej množiny MGT detektorov o personalizovane generované texty je prínosné, avšak optimálna voľba obohacujúceho datasetu závisí od cieľovej detekčnej domény. Pre maximalizáciu medzi-doménovej robustnosti je odporúčané kombinovať vysokú diverzitu štýlov, foriem, parametrov generovania, ako aj samotných generátorov (dostupné v datasete PerComm) s doménovo-relevantnými (priamo cielené na sociálne siete) personalizovanými vzorkami (dostupné v datasete PerQ). Naopak, ak je cieľom prevažne detekcia personalizovaných dezinformácií v doméne sociálnych sietí s vysokou seba-istotou detektorov, použitie celého datasetu PerQ (s najväčším počtom vzoriek textov sociálnych sietí) prináša najspoľahlivejšie výsledky. Tieto zistenia zároveň naznačujú, že vyššia miera personalizácie (najmä pri detailnej špecifikácii cieľovej skupiny) naďalej predstavuje čiastočnú výzvu pre spoľahlivosť detektorov, a to naprieč všetkými testovanými konfiguráciami, čo poukazuje na potrebu ďalšieho výskumu v tejto oblasti.

4 Referencie

[1] Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopál, Katarína Marcinčinová, and Matúš Mesarčík. 2025. Evaluation of LLM Vulnerabilities to Being Misused for Personalized Disinformation Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 780–797, Vienna, Austria. Association for Computational Linguistics.

[2] Dominik Macko and Andrew Pulver. 2025. PerQ: Efficient Evaluation of Multilingual Text Personalization Quality. arXiv preprint 2509.25903.

[3] Dominik Macko. 2026. Evaluation of Multilingual LLMs Personalized Text Generation Capabilities Targeting Groups and Social-Media Platforms. arXiv preprint 2601.03752.