CEAID: Benchmark of Multilingual Machine-Generated Text Detection Methods for Central European Languages

Dominik Macko and Jakub Kopal
Kempelen Institute of Intelligent Technologies
dominik.macko@kinit.sk, jakub.kopal@kinit.sk

Abstract

Machine-generated text detection, as an important task, is predominantly focused on English in research. This makes the existing detectors almost unusable for non-English languages, relying purely on cross-lingual transferability. There exist only a few works focused on any of Central European languages, leaving the transferability towards these languages rather unexplored. We fill this gap by providing the first benchmark of detection methods focused on this region, while also providing comparison of train-languages combinations to identify the best performing ones. We focus on multi-domain, multi-generator, and multilingual evaluation, pinpointing the differences of individual aspects, as well as adversarial robustness of detection methods. Supervised finetuned detectors in the Central European languages are found the most performant in these languages as well as the most resistant against obfuscation.

1 Introduction

Large language models (LLMs) are able to generate texts in various languages, hardly distinguishable for humans from authentic human-written texts. However, automated detection of such texts is mostly researched for English only (or other high-resource languages, such as Spanish or Russian), leaving some languages unprotected from massive spread of AI-generated content for malicious purposes (e.g., disinformation, spam, frauds, plagiarism). Such languages are often left to rely on cross-lingual transfer of monolingual detectors, which can have severely degraded performance.

It is important to explore the effect of such cross-lingual transferability to such languages and possibilities of introduction of multilingual detectors with an involvement of these languages in the detectors training or finetuning process. To the best of our knowledge, there is no study available focused on machine-generated text (MGT) detection in the languages of Central European region (see Figure 1), which is our region of interest due to project needs, especially focused on cross-lingual transferability. Furthermore, the combination of train languages has not been systematically studied in the MGT detection yet, which can alleviate a problem of missing data, especially in low-resource languages.

Figure 1: Central European region as defined by Bideleux and Jeffries, 2007.
source:https://en.wikipedia.org/wiki/Central_Europe#/media/File:Central-Europe-SwanseaUniv.png

Our study is specifically focused on answering the following research questions. ***RQ1:*** *Are there differences in finetuned detection methods performance based on combination of train languages?* If so, which combination of train languages makes the detectors the most generalizable to the other languages of the Central European region? ***RQ2:*** *Which category of detection methods is most suitable for Central European languages?* Are there differences among the MGT generation models? ***RQ3:*** *Which detection methods are most robust against obfuscation in Central European languages?* Are there differences in such robustness among the selected languages?

The contributions of our work can be summarized as[1]:

[1]For replication possibilities, all source code and data are released for non-commercial research purpose at https://github.com/kinit-sk/CEAID.

Findings of the Association for Computational Linguistics: ACL 2026, pages 29177–29190

- **the first comprehensive** (multilingual, multi-domain, multi-generator) **benchmark of MGT** detection methods focused on **Central European region**, reflecting cultural context of this specific region,
- the evaluation and **comparison of differences** in performance of MGT detection methods between **news and social-media domains**, reflecting different lengths and styles (e.g., formality, grammar, emoticons) of texts of these two domains,
- the evaluation of **train language combination effect** on generalizability and adversarial robustness of MGT detection methods, identifying the most suitable combination of train languages for the Central European region,
- introduction of a bunch of **language-specific MGT detection methods** for under-researched languages, which are usually out-of-focus of the mainstream MGT research relying purely on cross-lingual transferability of detectors (i.e., degraded performance).

2 Related Work

The related works are reviewed in three groups. First, the studies of applying MGT detection to non-English languages are summarized with interesting observations. Second, the existing MGT detection shared tasks are overviewed that are focusing not purely on English. Lastly, the existing benchmark datasets are summarized that focus on multiple languages, especially those that could be utilized in cross-lingual study.

2.1 Non-English MGT Detection

A study covering four languages (English, French, German, Spanish) of (South-)West European region shows that statistical features primarily developed for English MGT detection can be used in other languages as well (Schaaff et al., 2023). The detection has been however realized in monolingual way, i.e. testing on a single language that was included in training. Similarly, a study (Üyük et al., 2024), focused on English, Turkish, Hungarian, and Persian, has used a machine-learning classifiers on top of the TF-IDF features for monolingual detection.

Another study, focused on academic integrity, tested multiple existing directly usable (i.e., without training) detectors to examine their performance in machine translated texts (Weber-Wulff et al., 2023). The human written texts in Bosnian, Czech, German, Latvian, Slovak, Spanish, and Swedish, have been translated to English for the detection. They revealed that machine translated text had higher false positive rate than the text written by humans directly in English. It means that machine translation is not usable for MGT detection, and detection directly in non-English languages must be used. This is further supported by similar studies (Šigut and Foltỳnek, 2023; Bohacek, 2023), where the authors tested detection directly in Czech and Slovak and compared with translation to English. The translation negatively affected the performance of the used multilingual detector. They further identified that ChatGPT-4 was harder to detect than ChatGPT-3.5, pinpointing the need to conduct evaluation studies with most modern set of generators.

Most of the above mentioned studies have used ChatGPT (or its variants) as the only MGT generator (Üyük et al., 2024 used 5 different LLMs). Therefore, the generalizability of their conclusions is questionable. The study on detection of generated German texts (Irrgang et al., 2024) pin-pointed a worse cross-generator transferability of the detection performance, requiring robust detection methods. The study on detection of LLM-generated emails in Polish (Gryka et al., 2024) has identified that detection in underrepresented language in models training is inferior. The study on Bulgarian social-media texts (Temnikova et al., 2023) has shown that finetuned detectors on texts in that particular language can boost the performance significantly.

2.2 MGT Detection Shared Tasks

To drive the research direction into particular languages, there have recently been multiple MGT detection shared tasks focused monolingually on Russian at RuATD 2022 (Shamardina et al., 2022), Spanish at AuTexTification 2023 (Sarvazyan et al., 2023), Dutch at CLIN33 (Fivez et al., 2024). The IberAuTexTification 2024 (Sarvazyan et al., 2024) shared task was focused on multidomain detection in six languages (English, Spanish, Portuguese, Catalan, Basque, and Galician), especially targeting Iberian peninsula. It represents a regional focus for development of MGT detection methods, especially for low-resource languages; however, the detection in these languages has been executed individually (i.e., also in monolingual way). Thus,

cross-lingual transferability aspects have not been examined.

Recent multilingual shared tasks include SemEval-2024 Task 8 (Wang et al., 2024a) and GenAI Content Detection Task 1 (Wang et al., 2025), both of them representing multi-generator, multi-domain, and multilingual MGT detection challenge. However, in both of them, there is inconsistency in coverage of combinations of generators and domains across languages; thus any comparison among performances between languages is inherently biased. Furthermore, cross-lingual aspects are evaluated only towards few unseen languages during training; leaving the effect of cross-lingual transfer from particular languages unexplored.

2.3 Multilingual MGT Detection Benchmarks

The MultiSocial (Macko et al., 2025a) benchmark is focused on comparison of performance of MGT detection methods for social-media texts of 5 platforms in 22 languages. For generation of MGT, it uses 7 modern LLMs that 3-times paraphrase the original human-written texts. The number of samples per each platform and per each language is however not consistent, making the cross-lingual transferability experiments rather limited. The M4GT-Bench (Wang et al., 2024b) dataset has been used in the above-mentioned multilingual shared tasks. Inconsistency between per-language domain and generation settings makes the cross-lingual experiments inherently biased. Another benchmark focused on multilingual news articles, called MULTITuDE (Macko et al., 2023), covers 11 languages; however, only 3 of them contain training samples (English, Russian, and Spanish). The authors focus on cross-lingual transferability of detectors trained on these three languages towards the others in the test set. It is further extended to evaluation of adversarial robustness against 10 authorship obfuscation methods (Macko et al., 2024). It has identified a homoglyph-based obfuscation especially successful to evade the detection in multilingual settings. The RAID benchmark (Dugan et al., 2024), including RAID-extra part containing besides English also Python code and news articles in German and Czech, is focused on variable generation setting (decoding strategy) and 11 adversarial attacks. Such data are suited for evaluation of detectors robustness; however, language limitation disqualifies it from cross-lingual experiments.

3 Methodology

In order to provide the answers for our research questions, stated in Section 1, we need to craft a suitable dataset for the experiments with a good selection of languages, select suitable multilingual MGT detection methods of different categories for comparison, and come up with proper settings for rigorous evaluation enabling generalizability of the conclusions. All of these are reflected in the following subsections, while limiting the scope of the study to keep it feasible and prevent waste of resources (especially time and computational requirements).

3.1 Dataset

Our goal was to ensure consistency among languages, enabling rigorous cross-lingual evaluation. For generalization of our observations, we included multiple generators and multiple domains. Specifically, we have selected the Central European languages from MULTITuDE_v3 (Macko et al., 2025b) (domain of news articles) and MultiSocial (Macko et al., 2025a) (domain of social-media texts from 5 platforms), having at least 200 samples (for test split) per each domain and class (human vs. machine). It resulted into selection of 7 languages, namely Croatian, Czech, German, Hungarian, Polish, Slovak, and Slovenian, together covering 3 language-family branches of Germanic, Slavic, and Uralic (all using Latin writing script). 2 languages, Slovak and Slovenian, are not used for training due to not having enough samples from the social-media domain available. The overview of the resulted sample counts per each language, domain, and split of the selected dataset is provided in Table 1. The MGTs are generated by 8 LLMs in total, 6 of which are the same across the two domains (Aya-101, GPT-3.5-Turbo-0125, Mistral-7B-Instruct-v0.2, OPT-IML-Max-30B, v5-Eagle-7B-HF, and Vicuna-13B), one is only in news domain (Llama-2-70B-chat-hf), and one is only in

Domains →	**News**		**Social media**		**All**	
Language ↓	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
cs (Czech)	7734	2328	11041	6073	18775	8401
de (German)	7764	2322	21038	9497	28802	11819
hr (Croatian)	7819	2348	14475	5993	22294	8341
hu (Hungarian)	7791	2350	14492	5957	22283	8307
pl (Polish)	7818	2336	16687	6971	24505	9307
sk (Slovak)	7664	2317	0	2026	7664	4343
sl (Slovenian)	7845	2354	0	3058	7845	5412
Total	54435	16355	77733	39575	132168	55930

Table 1: Overview of the selected dataset sample counts.

social-media domain (Gemini). Since the generators are consistent among the languages (uniformly distributed sample counts), we do not see a problem to include all of them in regard to our research questions (not only the intersection).

3.2 Detectors

We follow the previous cross-lingual benchmark studies (Macko et al., 2023, 2025a) to select 3 categories of MGT detection methods: *statistical* (zero-shot), pretrained (directly applicable), and finetuned (trained on train split).

As statistical detectors, we are using **Binoculars** (Hans et al., 2024), **Fast-DetectGPT** (Bao et al., 2023), and **LLM-Deviation** (Wu and Xiang, 2023), all of them based on multilingual mGTP LLM (Shliazhko et al., 2024) (as both the reference and sampling models of Fast-DetectGPT or the observer and performer models of Binoculars).

As pretrained detectors, we have selected **ChatGPT-detector-RoBERTa-Chinese** (Guo et al., 2023), **Longformer Detector** (Li et al., 2024), **BLOOMZ-3B-mixed-detector** (Sivesind and Winje, 2023). All of the mentioned detectors have been selected due to performing well in each category in the MultiSocial (Macko et al., 2025a) study or its cross-domain evaluation. Analogously to the above mentioned study, we have used the published source code of the IMGTB framework (Spiegel and Macko, 2024) to run these detectors.

As finetuned detectors, we have selected **mDeBERTa-v3-base** (He et al., 2022) and **XLM-RoBERTa-base** (Conneau et al., 2020) (as multilingual baselines), **Llama-3.2-3B** (Meta, 2024) (as a newer and smaller version of the best performing MultiSocial-finetuned model), and **Gemma-2-2B** (Team, 2024) (as a smaller version of the best performing model in out-of-distribution evaluation of Macko et al., 2025c).

3.3 Settings

For strong representativeness of our results and generalization of conclusions, we have carefully designed the experiments. Targeting RQ1, we have selected 4 models for finetuning of diverse architectures (encoder vs decoder) and sizes (number of parameters ranging from 0.3B to 3B). For RQ2, we have used 3 categories of detection methods, each containing at least 3 methods (based on existing studies). For RQ3, we have used two prominent obfuscation methods (paraphrasing for its usability and homoglyphs for their effectiveness) and 10 diverse detectors.

As the primary evaluation metric, we use **AUC ROC** (area under curve of the receiver operating characteristic) as a classification-threshold independent metric. We are also providing **TPR @ 5% FPR** in the appendix for a deeper analysis, representing true positive rate (TPR) using the classification thresholds calibrated (based on ROC curve) to reach 5% of false positive rate (FPR), reflecting an expected performance in the wild.

To further ensure consistency between the two domains and between the two classes (human and machine), in each experiment, we have used pseudo-random sub-sampling of training samples to the highest possible number to achieve the perfect balance. This number is the lowest count of human-written texts out of each domain for each language, being 986. It resulted into training set of 3944 samples per language (986×2 [human vs machine] $\times 2$ [news vs social-media]). This number of train samples is kept the same in all experiments (monolingual training as well as multilingual training with even portion of each train language). For testing and unbiased comparison, we have further sub-sampled 250 samples from the test split per each class, domain, and language, and 200 samples per each generator, domain and language for comparison of detection performance per generators. These numbers are reflecting the lowest count of test samples for any combination to achieve perfect balance.

To save computational resources for adversarial robustness evaluation, we have sub-sampled 100 samples per each class, domain, and language from test set (i.e., a subset of 2800 samples) to be further paraphrased by DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025) (as a most modern highly-performant multilingual LLM) and obfuscated by the generic HomoglyphAttack (Macko et al., 2024). The comparison between above-mentioned (bigger) test set evaluation and this smaller subset ensures its representativeness. From the finetuned models (31 per each of 4 base models), we have selected for evaluation of adversarial robustness only de-hu-pl trained versions of each base model (as representatives of the three different language-family branches and achieving one of the highest performances).

4 Results

The results are divided into three parts based on the addressed research questions.

4.1 Training Languages Evaluation

To evaluate effect of a combination of training languages in the model finetuning on MGT detection, we provide the mean performance (AUC ROC) per each train languages combination, which is averaged across the four base models used for finetuning. These mean values along with standard deviations per each test language are reported in Table 2. The results are sorted based on the mean performance for all test languages combined (All).

There are differences (some of which are statistically significant) in the MGT performance among the combinations of train languages for finetuning. The results indicate that German and Polish are very important to include in finetuning (*de* is in 9 of the top 10 performing language combinations and *pl* is in 8 of the top 10). However, it is also important to combine at least two train languages, as all of the single-train-language versions of the finetuned models ranked in the bottom 10. In 8 of top 10 versions, at least three languages are combined. A paired t-test ($\alpha = 0.05$) identified approximately one third of differences between these train-languages combinations as statistically significant (e.g., the top-performing vs. worst-performing combination among them).

It seems that the generalization to Slovenian (i.e., cross-lingual transfer due to Slovenian not present in the training) is the most difficult, requiring Polish or Czech to be present in the finetuning to reach the best performance. Surprisingly, **Croatian has clearly the worst cross-lingual transferability to Slovenian**, although being the language of the neighboring country of Slovenia and being from the same language-family branch.

There is a small variability across the base models, where the standard deviation is reaching up to 6% (in most cases under 1%). The difference between the top-performing combination of train languages and the worst-performing combination is about 3%. Most of the differences are not statistically significant. Therefore, all of these indicate the performance is quite stable. It seems that it does not matter as much on the base model selected for finetuning as to include a combination of at least two languages for training (ideally from different language families).

4.2 Detectors Categories Comparison

We have compared the performance of three different categories of MGT detectors, namely statistical, pretrained, and finetuned. The comparison of the detectors is provided in Table 3. In the table, we are showing only the best performing combination of train languages for each base model for better readability and space limitation (the worst performing combination is not significantly lower as can be seen in Table 2). The comparison of the selected MGT detectors shows that the **finetuned detectors are consistently the best performing category across all languages**, followed by statistical methods. The worst performing are pretrained detectors, for which this Central European set of languages might be too out-of-distribution. The finetuned detectors achieve by more than 10% higher performance than the statistical detectors. The difference between statistical and pretrained categories is not that high, for some languages some pretrained detectors outperformed the worst of statistical detectors (LLM-Deviation).

When looking at the results per each MGT generation model, provided in Table 4, we can observe differences especially for statistical detectors. They are best at detecting Llama-2 generated texts (around 0.9 of AUC ROC); on the other hand, Mistral or OPT-IML data are the hardest for them (around 0.6 of AUC ROC). The further analysis is needed to explore this phenomenon. We can speculate that this is due to worse data quality of MGTs generated by these two models in Central European languages, but there can be a relationship between the mGPT model used as a base model for statistical detectors and the generators (either with the good performance or those two with the low performance).

To compare the performance of MGT detectors for each of the two domains (news and social media), we are providing the per-language comparison for each domain separately in Table 5, analogously to Table 3. We can observe that the **finetuned detectors dominate in both domains**, while news articles (longer and more formal texts) are a bit easier for the finetuned detectors. Interestingly, none of the train-languages combination of the best-performing finetuned models is the same across the two domains; therefore, there is none clearly dominating combination. Furthermore, for mDeBERTa base model, a single-train language (Czech for News, German for Social media) achieved the

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

Table 2: Per-test-language comparison of performance (AUC ROC averaged across the finetuned base models) of finetuned MGT detectors based on combination of train languages. Bold represents the highest value per each test language.

Category	Detector	All	cs	de	hr	hu	pl	sk	sl
F	Llama-3.2-3B (de-hu-pl)	**0.9758**	**0.9886**	**0.9739**	**0.9829**	**0.9851**	**0.9765**	**0.9779**	**0.9638**
F	mDeBERTa-v3-base (cs-de-hr-pl)	0.9739	0.9835	0.9731	0.9789	0.9821	0.9727	0.9693	0.9624
F	Gemma-2-2B (cs-de-pl)	0.9660	0.9837	0.9697	0.9546	0.9750	0.9620	0.9765	0.9434
F	XLM-RoBERTa-base (cs-de-hr-hu-pl)	0.9621	0.9778	0.9484	0.9744	0.9748	0.9541	0.9606	0.9497
S	Fast-DetectGPT	0.7904	0.7943	0.8074	0.8228	0.7587	0.7830	0.7829	0.8133
S	Binoculars	0.7675	0.7811	0.7924	0.8000	0.7517	0.7681	0.7496	0.7650
S	LLM-Deviation	0.6887	0.7543	0.6666	0.7258	0.6855	0.6991	0.7141	0.7337
P	BLOOMZ-3B-mixed-detector	0.6740	0.6769	0.6945	0.6690	0.6836	0.6752	0.7292	0.5997
P	ChatGPT-detector-RoBERTa-Chinese	0.6492	0.6045	0.7055	0.6442	0.7238	0.6361	0.6885	0.6629
P	Detection-Longformer	0.5629	0.5687	0.4860	0.6519	0.6319	0.5900	0.4636	0.5654

Table 3: Per-test-language comparison of performance (AUC ROC) of categories of MGT detectors (S – statistical, P – pretrained, F – finetuned). For readability, the finetuned category includes only the best performing combination of train languages of each base model. Bold represents the highest value per each test language.

best performance. Polish and Slovenian social-media texts are the most difficult for detection. In other categories, LLM-Deviation and BLOOMZ-3B-mixed-detector are the worst in the news domain, while being the best in the social-media domain. Detection-Longformer has similarly the better performance in news, while performing worse than random classifier in the social-media domain. For pretrained detectors, it definitely depends on the domain of their pretraining data. But, the case is different for LLM-Deviation, since it is zero-shot statistical detection metric (i.e., without training) and this effect is not consistent among languages (most significant in Hungarian and Polish). There is a moderate Pearson correlation of statistical detectors with the text length (much shorter texts in social media as analyzed in Table 8). The finetuned and pretrained detectors have weak or no linear relationship between machine-class probability and text length (up to 0.3 of Pearson ρ).

Category	Detector	All	Llama-2-70B-Chat-HF	Mistral-7B-Instruct-v0.2	Aya-101	Gemini	GPT-3.5-Turbo-0125	OPT-IML-Max-30B	v5-Eagle-7B-HF	Vicuna-13B
F	Llama-3.2-3B (de-hu-pl)	**0.9749**	0.9904	0.9688	0.9655	**0.9754**	**0.9828**	**0.9632**	0.9822	0.9789
F	mDeBERTa-v3-base (cs-de-hr-hu-pl)	0.9734	**0.9935**	**0.9704**	**0.9662**	0.9706	0.9797	0.9505	**0.9848**	**0.9801**
F	Gemma-2-2B (cs-de-pl)	0.9627	0.9777	0.9555	0.9612	0.9585	0.9717	0.9452	0.9722	0.9653
F	XLM-RoBERTa-base (cs-de-hr-hu-pl)	0.9613	0.9821	0.9430	0.9526	0.9626	0.9745	0.9403	0.9732	0.9732
S	Fast-DetectGPT	0.7830	0.9667	0.6305	0.8292	0.7920	0.7986	0.6377	0.8729	0.8325
S	Binoculars	0.7603	0.9555	0.6211	0.7955	0.7889	0.7756	0.5985	0.8518	0.8075
S	LLM-Deviation	0.6843	0.8873	0.6233	0.6807	0.6584	0.6861	0.5610	0.7570	0.7088
P	BLOOMZ-3B-mixed-detector	0.6730	0.5410	0.6302	0.6655	0.6271	0.7694	0.6697	0.7000	0.6923
P	ChatGPT-detector-RoBERTa-Chinese	0.6423	0.7949	0.6784	0.6090	0.6906	0.5882	0.5667	0.6403	0.6711
P	Detection-Longformer	0.5615	0.6601	0.6045	0.5209	0.5045	0.4613	0.5453	0.6283	0.5879

Table 4: Per-generator comparison of performance (AUC ROC) of categories of MGT detectors (S – statistical, P – pretrained, F – finetuned). For readability, the finetuned category includes only the best performing combination of train languages of each base model. Bold represents the highest value per each MGT generator.

Domain	Detector	All	cs	de	hr	hu	pl	sk	sl
News	Llama-3.2-3B (cs-hr-hu)	**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 (cs-de-pl)	0.9911	0.9966	0.9912	0.9827	0.9882	0.9878	0.9978	0.9894
	XLM-RoBERTa-base (cs-de-hr-hu-pl)	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-hu-pl)	**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 (cs-de-hr-hu-pl)	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

Table 5: Per-test-language comparison of performance (AUC ROC) of the selected MGT detectors for each domain. For readability, the finetuned category includes only the best performing combination of train languages of each base model (for each domain). Bold represents the highest value per each test language and each domain.

4.3 Adversarial Robustness Evaluation

We have evaluated adversarial robustness of the MGT detectors against obfuscation by paraphrasing and by homoglyph attack. The results, provided in Table 6 and Table 7, indicate that **finetuned detectors are the most robust towards obfuscation** (especially bigger ones). With a sole exception of Detection-Longformer (where the performance is actually increased), the **homoglyph-based obfuscation decreases the performance significantly more than paraphrasing** (in all languages). Detection-Longformer is more susceptible to paraphrasing. The other two pretrained detectors and finetuned detectors are mostly immune to this kind of obfuscation across all the languages. Statistical detectors are confused by both obfuscation methods, while homoglyph attack can decrease the AUC ROC performance by 95% (in case of Fast-DetectGPT for German texts).

5 Conclusions

This benchmark study, focused on a set of Central European languages, brought several insights in the detection of machine-generated texts for a bunch of under-researched languages. The comparison of the effect of language combinations on detectors

Detector	Subset	All	cs	de	hr	hu	pl	sk	sl
Llama-3.2-3B	original	0.9739	0.9787	0.9717	0.9881	0.9825	0.9808	0.9703	0.9654
	paraphrased	**0.9851**	**0.9921**	**0.9785**	**0.9966**	**0.9869**	**0.9867**	**0.9845**	**0.9827**
	homoglyph	0.9354	0.9630	0.9282	0.9726	0.9507	0.9347	0.9278	0.9157
mDeBERTa-v3-base	original	0.9720	0.9769	0.9734	0.9674	0.9816	0.9757	0.9721	0.9636
	paraphrased	**0.9825**	**0.9881**	**0.9813**	**0.9782**	**0.9869**	**0.9828**	**0.9888**	**0.9744**
	homoglyph	0.7749	0.8519	0.6736	0.8080	0.8215	0.7367	0.8170	0.7024
Gemma-2-2B	original	0.9588	0.9769	0.9647	0.9634	0.9783	0.9752	0.9677	0.9484
	paraphrased	**0.9779**	**0.9889**	**0.9750**	**0.9908**	**0.9809**	**0.9826**	**0.9832**	**0.9741**
	homoglyph	0.8962	0.9377	0.9326	0.9060	0.9530	0.9115	0.9011	0.8589
XLM-RoBERTa-base	original	0.9540	0.9633	0.9474	0.9627	0.9736	0.9611	0.9434	0.9322
	paraphrased	**0.9665**	**0.9769**	**0.9536**	**0.9774**	**0.9805**	**0.9629**	**0.9687**	**0.9536**
	homoglyph	0.5218	0.6393	0.3788	0.5510	0.6389	0.5300	0.4809	0.4128
Fast-DetectGPT	original	**0.8000**	**0.8045**	**0.8122**	**0.8049**	**0.8154**	**0.8096**	**0.7658**	**0.8112**
	paraphrased	0.7074	0.7657	0.7921	0.7229	0.5298	0.6965	0.7343	0.7394
	homoglyph	0.0815	0.1093	0.0403	0.0542	0.0808	0.0857	0.1232	0.0750
Binoculars	original	**0.7758**	**0.7894**	**0.7917**	**0.7737**	**0.8080**	**0.7993**	**0.7376**	**0.7601**
	paraphrased	0.7098	0.7683	0.7855	0.7171	0.5951	0.7188	0.7219	0.7121
	homoglyph	0.2711	0.2951	0.2338	0.2313	0.3252	0.2866	0.2996	0.2214
LLM-Deviation	original	**0.6968**	**0.7473**	0.6958	**0.7030**	**0.7288**	**0.7208**	0.7114	**0.7035**
	paraphrased	0.6775	0.7322	**0.7044**	0.6964	0.5978	0.6622	**0.7158**	0.6881
	homoglyph	0.3614	0.4219	0.2454	0.3539	0.4082	0.3527	0.3466	0.3136
BLOOMz-3B-mixed-detector	original	0.6778	0.6565	0.6707	0.7015	0.6923	0.6879	0.7095	0.6450
	paraphrased	**0.8487**	**0.8526**	**0.8379**	**0.8527**	**0.8337**	**0.8754**	**0.8904**	**0.8275**
	homoglyph	0.4240	0.4333	0.4203	0.4280	0.4281	0.4369	0.4562	0.3627
ChatGPT-detector-RoBERTa-Chinese	original	0.6497	**0.6228**	**0.6915**	0.6473	0.7228	**0.6294**	0.6655	**0.6769**
	paraphrased	**0.6515**	0.6008	0.6829	**0.6741**	**0.7408**	0.6196	**0.6955**	0.6278
	homoglyph	0.5547	0.5127	0.5404	0.5184	0.6221	0.5453	0.5516	0.5045
Detection-Longformer	original	0.5616	0.6032	**0.5010**	0.6065	0.6179	0.6097	0.4760	0.5391
	paraphrased	0.5107	0.5247	0.4216	0.5887	0.5287	0.5683	0.4253	0.5246
	homoglyph	**0.6196**	**0.6319**	0.4883	**0.6870**	**0.6492**	**0.6504**	**0.5537**	**0.6804**

Table 6: Per-test-language comparison of performance (AUC ROC) of the selected MGT detectors on original and adversarial data. Bold represents the highest value per each test language and each detector (in regard to original, paraphrased, and homoglyph subset).

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**

Table 7: Per-test-language evaluation of adversarial robustness of the selected MGT detectors as a difference in performance (AUC ROC) on the obfuscated adversarial data and on the original data. Bold represents the highest value per each test language and each detector (in regard to individual obfuscated subsets).

finetuning revealed small differences in generalization towards other Central European languages. Out of the compared statistical, pretrained, and finetuned categories of MGT detectors, the last one is performing significantly better across all the tested languages. The finetuned detectors are also the most robust against paraphrasing and homoglyph-based obfuscation, making them the most suitable for Central European languages. However, until now there have been only few models available in some of these languages. This further signifies the need to perform research also in the languages left usually out-of-focus of the mainstream.

Limitations

Although this study covers 7 languages, 2 domains, 8 LLMs, 2 authorship obfuscation techniques, and 10 MGT detection methods, the results might still be biased based on some specific aspects of the data, affecting the generalizability of the conclusions to out-of-distribution data (e.g., other languages, other LLM generators). For example, the selected languages cover 3 language-family branches, but 5 of the languages are Slavic. This is, however, specific bias of the Central European geographic area, which was the aim of this study. The cross-lingual transferability is also biased based on pre-training data of individual detection base models (although we focused on multilingual versions). We have done a manual hyper-parameters optimization of the detectors' finetuning process; however, we have covered just a small set of options, where a further tuning might increase the performance and generalizability of the individual detectors.

Ethics Statement

Our work is focused on evaluation and comparison (i.e., benchmarking) of the MGT detection methods on 7 languages of Central European region, bringing important insights in these under-researched languages. We use the existing datasets in our work in accordance with their intended use and licenses (for research purpose only). As a part of our work, we are not re-sharing any existing data or publishing any new dataset. For the research replicability and validation purposes, we are publishing the preprocessing, training, and evaluation source codes (for research purposes only). The existing artifacts used in this work have been properly cited and used according their licenses and intended use. We have also checked and followed licensing and terms of use of the used LLMs. AI assistants have not been used for conducting research in any other way than already described in the paper (text obfuscation, finetuning and detection of MGTs).

Acknowledgments

Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00059.

Computational resources. Part of the research results was obtained using the computational resources procured in the national project *National competence centre for high performance computing* (project code: 311070AKF2) funded by European Regional Development Fund, EU Structural Funds Informatization of Society, Operational Program Integrated Infrastructure.

References

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Robert Bideleux and Ian Jeffries. 2007. *A history of Eastern Europe: Crisis and change*. Routledge.

Matyas Bohacek. 2023. The unseen A+ student: Navigating the impact of large language models in the classroom. In *Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Pieter Fivez, Walter Daelemans, Tim Van de Cruys, Yury Kashnitsky, Savvas Chamezopoulos, Hadi Mohammadi, Anastasia Giachanou, Ayoub Bagheri, Wessel Poelman, Juraj Vladika, Esther Ploeger, Johannes Bjerva, Florian Matthes, and Hans van Halteren. 2024. The CLIN33 shared task on the detection of text generated by large language models. *Computational Linguistics in the Netherlands Journal*, 13:233–259.

Paweł Gryka, Kacper Gradoń, Marek Kozłowski, Miłosz Kutyła, and Artur Janicki. 2024. Detection of ai-generated emails - a case study. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ARES '24, New York, NY, USA. Association for Computing Machinery.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Verena Irrgang, Veronika Solopova, Steffen Zeiler, Robert M. Nickel, and Dorothea Kolossa. 2024. Features and detectability of German texts generated with large language models. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 264–280, Vienna, Austria. Association for Computational Linguistics.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025a. MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 727–752, Vienna, Austria. Association for Computational Linguistics.

Dominik Macko, Jakub Kopal, Robert Moro, and Ivan Srba. 2025b. MULTITuDEv3.

Dominik Macko, Robert Moro, and Ivan Srba. 2025c. Increasing the robustness of the fine-tuned multilingual machine-generated text detectors. *Preprint*, arXiv:2503.15128.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9960–9987, Singapore. Association for Computational Linguistics.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason S Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024. Authorship obfuscation in multilingual machine-generated text detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6348–6368, Miami, Florida, USA. Association for Computational Linguistics.

AI Meta. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *Procesamiento del Lenguaje Natural*, 71:275–288.

Areg Mikael Sarvazyan, José Ángel González, Francisco Rangel, Paolo Rosso, and Marc Franco-Salvador. 2024. Overview of iberautextification at IberLEF 2024: Detection and attribution of machine-generated text on languages of the Iberian peninsula. *Procesamiento del Lenguaje Natural*, (73):421–434.

Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2023. Classification of human- and AI-generated texts for English, French, German, and Spanish. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP*

2023), pages 1–10, Online. Association for Computational Linguistics.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the RuATD shared task 2022 on artificial text detection in Russian. In *Computational Linguistics and Intellectual Technologies*. RSUH.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Petr Šigut and Tomáš Foltýnek. 2023. Can we detect ChatGPT-generated texts in Czech and Slovak languages? *RASLAN 2023 Recent Advances in Slavonic Natural Language Processing*, page 35.

Nicolai Thorer Sivesind and Andreas Bentzen Winje. 2023. Machine-generated text-detection by fine-tuning of language models.

Michal Spiegel and Dominik Macko. 2024. IMGTB: A framework for machine-generated text detection benchmarking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 172–179, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team. 2024. Gemma.

Irina Temnikova, Iva Marinova, Silvia Gargova, Ruslana Margova, and Ivan Koychev. 2023. Looking for traces of textual deepfakes in Bulgarian on social media. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1151–1161, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Cem Üyük, Danica Rovó, Shaghayeghkolli Shaghayeghkolli, Rabia Varol, Georg Groh, and Daryna Dementieva. 2024. Crafting tomorrow's headlines: Neural news generation and detection in English, Turkish, Hungarian, and Persian. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 271–307, Miami, Florida, USA. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024a. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, and 7 others. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 244–261, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):1–39.

Zhendong Wu and Hui Xiang. 2023. MFD: Multi-feature detection of LLM-generated text. *PREPRINT (Version 1) available at Research Square*.

A Computational Resources

For finetuning and inference of machine-generated text detection models, we have used 1x A100 40GB, consuming about 2000 GPU-hours. For executing authorship obfuscation in machine-generated texts and subsequent text-quality analysis, we have used 1x A100 40GB, consuming about 70 GPU-hours. For other tasks, we have not used GPU acceleration.

For finetuning, we have used python transformers library and using half-precision LoRA parameter-efficient finetuning process with learning rate of 3e-4, maximum context length of 512 input tokens, 8 epochs, and micro-batch size of 32 (which has been manually decreased up to 1 for bigger models to fit the utilized GPU). The hyper-parameters have been manually optimized for models to successfully learn the task. For replication purpose, the random seed has been set to the value of 1337. More details can be found in the enclosed source code. For paraphrasing by DeepSeek-R1-Distill-Qwen-32B, we have used a simple paraphrase prompt of "Paraphrase and polish the following text. Keep it in the original language and correct the grammar and stylistic errors.

Generator	News		Social media	
	CC	WC	CC	WC
Human	795.99 (±269.48)	115.06 (±38.77)	66.86 (±92.36)	9.09 (±11.16)
Llama-2-70B-Chat-HF	986.08 (±368.97)	147.24 (±57.12)	-	-
Mistral-7B-Instruct-v0.2	965.22 (±305.23)	137.08 (±44.46)	107.82 (±74.68)	15.03 (±9.57)
Aya-101	1069.99 (±420.08)	163.41 (±65.00)	60.12 (±60.09)	9.65 (±8.73)
Gemini	-	-	435.09 (±365.40)	65.68 (±50.87)
GPT-3.5-Turbo-0125	742.37 (±260.78)	105.05 (±36.25)	100.38 (±99.01)	16.67 (±16.85)
OPT-IML-Max-30B	491.70 (±353.00)	74.47 (±42.75)	47.02 (±44.46)	7.43 (±6.57)
v5-Eagle-7B-HF	923.74 (±347.38)	135.26 (±51.95)	111.15 (±85.21)	16.87 (±12.35)
Vicuna-13B	840.92 (±343.39)	124.90 (±51.55)	93.15 (±77.87)	14.16 (±11.11)

Table 8: Dataset statistics of mean count (+- standard deviation) of characters (CC) and words (WC) per generator and per domain.

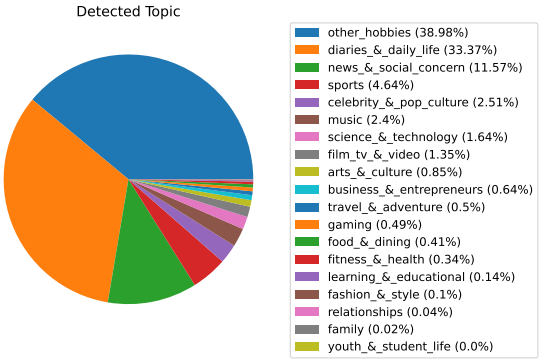

Figure 2: Detected topics in the selected dataset.

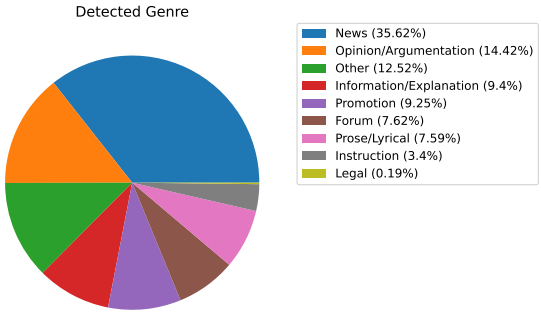

Figure 3: Detected genres in the selected dataset.

Then polish the generated text to appear more human written in the original language. Output just the final corrected and polished text." We have used 4-bit inference with min_new_tokens of 5, max_new_tokens of 32768, temperature of 0.6, repetition_penalty of 1.1, top_k of 50, and top_p of 0.95. Other text-generation parameters used the default values.

B Data Analysis

We have analyzed basic stylometric characteristics of the selected combined dataset in a form of character counts and word counts per each included generator and domain. The results of such analysis are summarized in Table 8. Aya generated the longest news articles and Gemini generated the longest social-media texts. On the other hand, OPT-IML generated the shortest texts in both domains.

Similarly to the MultiSocial study, we have used analysis of topics and genres based on existing available detectors. The overview of the results of the topic detector[2] (Antypas et al., 2022) is illustrated in Figure 2. The results overview of the multilingual text genre detector[3] (Kuzman et al., 2023) is illustrated in Figure 3. Although the texts are combined from news and social-media domains, there is variety of topics and genres in the data, making the results of this study representative (limiting the topical bias).

After application of the two authorship obfuscation techniques on subset of data, we have run a similarity analysis using various standard metrics, defined by Macko et al., 2024, to compare original and obfuscated texts (Table 9). It seems that paraphrasing significantly prolonged the texts, which might indicate an easier subsequent detectability. The semantic similarity indicated by BERTScore

[2] https://huggingface.co/cardiffnlp/tweet-topic-latest-multi

[3] https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier

	Paraphrased	Homoglyph
METEOR	0.477 (±0.23)	0.216 (±0.24)
BERTScore	0.824 (±0.09)	0.872 (±0.05)
ngram	0.429 (±0.20)	0.619 (±0.08)
TF	0.684 (±0.22)	0.311 (±0.25)
LD	1.489 (±16.15)	0.099 (±0.03)
CharLenDiff	1.988 (±16.18)	1.061 (±0.35)
LangCheck	17.64%	14.64%

Table 9: Similarity of obfuscated texts to the original.

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	**0.849**	0.036	0.936	0.019	0.844	0.080	0.893	0.020	0.922	0.038	0.825	0.054	0.863	0.041	0.275	0.337
de-pl	0.844	0.033	0.916	0.023	0.863	0.045	0.850	0.065	0.901	0.050	0.849	0.062	0.845	0.056	0.731	0.033
cs-de	0.840	0.037	0.937	0.019	**0.879**	0.039	0.855	0.051	0.893	0.047	0.762	0.086	0.855	0.045	**0.750**	0.067
cs-de-pl	0.837	0.037	0.932	0.021	0.843	0.056	0.820	0.086	0.909	0.032	0.837	0.057	0.858	0.052	0.626	0.258
de-hu-pl	0.833	0.046	0.920	0.039	0.840	0.066	0.846	0.059	0.929	0.044	0.829	0.074	0.849	0.039	0.746	0.074
cs-de-hr-pl	0.832	0.033	0.928	0.018	0.850	0.043	0.886	0.027	0.865	0.039	0.817	0.051	0.821	0.057	0.528	0.356
de-hr-pl	0.822	0.045	0.928	0.028	0.838	0.047	0.889	0.049	0.897	0.031	0.804	0.095	**0.886**	0.029	0.351	0.405
cs-hu-pl	0.813	0.061	**0.941**	0.033	0.572	0.387	0.868	0.038	**0.951**	0.020	0.845	0.061	0.820	0.101	0.650	0.100
de-hu	0.810	0.046	0.922	0.017	0.868	0.040	0.815	0.071	0.942	0.021	0.736	0.053	0.832	0.022	0.699	0.032
cs-de-hr	0.802	0.060	0.933	0.030	0.851	0.037	0.877	0.033	0.899	0.021	0.575	0.383	0.824	0.061	0.002	0.003
cs-hr-pl	0.801	0.045	0.934	0.031	0.521	0.369	0.890	0.014	0.906	0.035	0.837	0.054	0.821	0.061	0.336	0.394
cs-de-hu-pl	0.797	0.067	0.915	0.046	0.841	0.077	0.674	0.367	0.905	0.046	0.812	0.059	0.826	0.081	0.569	0.304
hu-pl	0.793	0.047	0.926	0.038	0.308	0.360	0.822	0.061	0.950	0.020	0.599	0.403	0.785	0.072	0.466	0.313
cs-de-hu	0.783	0.061	0.897	0.094	0.852	0.073	0.770	0.109	0.923	0.038	0.477	0.380	0.596	0.402	0.397	0.330
cs-hr-hu	0.782	0.104	0.916	0.068	0.742	0.162	0.867	0.046	0.926	0.070	0.689	0.149	0.785	0.098	0.263	0.307
cs	0.779	0.054	0.928	0.011	0.595	0.397	0.810	0.077	0.877	0.075	0.403	0.466	0.631	0.422	0.679	0.056
de	0.776	0.074	0.879	0.044	0.875	0.034	0.767	0.102	0.816	0.129	0.719	0.058	0.825	0.064	0.709	0.063
pl	0.772	0.119	0.907	0.084	0.427	0.493	0.849	0.062	0.895	0.068	**0.858**	0.038	0.820	0.127	0.710	0.083
cs-de-hr-hu	0.769	0.097	0.927	0.030	0.860	0.060	**0.903**	0.025	0.935	0.031	0.751	0.082	0.570	0.405	0.171	0.341
cs-pl	0.740	0.133	0.900	0.107	0.380	0.443	0.799	0.088	0.876	0.073	0.810	0.118	0.799	0.116	0.512	0.341
cs-hu	0.686	0.194	0.913	0.042	0.644	0.225	0.828	0.038	0.932	0.030	0.662	0.139	0.673	0.206	0.514	0.120
hu	0.664	0.113	0.855	0.082	0.649	0.165	0.745	0.120	0.933	0.046	0.654	0.189	0.690	0.090	0.582	0.143
de-hr-hu-pl	0.628	0.422	0.927	0.029	0.851	0.065	0.891	0.016	0.689	0.460	0.618	0.420	0.627	0.419	0.189	0.378
hr-hu	0.599	0.401	0.906	0.057	0.575	0.384	0.865	0.039	0.934	0.021	0.476	0.356	0.411	0.475	0.206	0.412
cs-hr-hu-pl	0.596	0.400	0.937	0.024	0.535	0.374	0.877	0.023	0.946	0.030	0.817	0.057	0.573	0.385	0.171	0.341
hr-hu-pl	0.594	0.398	0.931	0.039	0.349	0.410	0.643	0.430	0.944	0.022	0.825	0.075	0.567	0.392	0.172	0.339
de-hr	0.594	0.400	0.918	0.028	0.875	0.042	0.902	0.007	0.900	0.040	0.628	0.264	0.840	0.035	0.072	0.145
de-hr-hu	0.555	0.383	0.899	0.059	0.822	0.087	0.877	0.029	0.936	0.015	0.731	0.075	0.567	0.406	0.162	0.324
hr-pl	0.391	0.454	0.931	0.033	0.550	0.377	0.901	0.019	0.910	0.029	0.834	0.077	0.782	0.116	0.192	0.383
hr	0.345	0.403	0.893	0.055	0.535	0.366	0.877	0.044	0.843	0.100	0.438	0.347	0.446	0.442	0.068	0.135
cs-hr	0.168	0.335	0.914	0.039	0.372	0.442	0.854	0.040	0.885	0.050	0.399	0.462	0.589	0.404	0.186	0.372

Table 10: Per-test-language comparison of performance (TPR @ 5% FPR averaged across the finetuned base models) of finetuned MGT detectors based on combination of train languages. Bold represents the highest value per each test language.

seems to be high after both obfuscations. Language detection seems to be not very accurate in shorter social-media texts.

C Results Data

Table 10, Table 11, and Table 12 contain the performance comparison using the TPR @ 5% FPR metric, reflecting expected performance in the real world (where FPR must be minimized). These results indicate that even the finetuned detectors are far from perfect and the performance must be further tuned in the future.

Category	Detector	All	cs	de	hr	hu	pl	sk	sl
F	Llama-3.2-3B (pl)	**0.8954**	0.9480	**0.8780**	**0.9260**	**0.9500**	**0.8940**	**0.9300**	0.7520
F	mDeBERTa-v3-base (de-hr-hu-pl)	0.8749	0.9300	**0.8780**	0.8980	0.9180	0.8900	0.8400	**0.7560**
F	Gemma-2-2B (de-hr-hu-pl)	0.8731	**0.9500**	**0.8780**	0.9080	0.9380	0.8760	0.8720	0.0000
F	XLM-RoBERTa-base (cs-de)	0.8206	0.9280	0.8200	0.8520	0.8260	0.7340	0.8260	**0.7560**
S	Fast-DetectGPT	0.4089	0.3980	0.5120	0.4580	0.4280	0.4520	0.3720	0.4140
S	Binoculars	0.3871	0.4740	0.5040	0.5100	0.4240	0.4200	0.3520	0.4160
S	LLM-Deviation	0.1806	0.3880	0.2560	0.4300	0.2020	0.3040	0.2540	0.4280
P	BLOOMZ-3B-mixed-detector	0.1540	0.1840	0.2140	0.1500	0.1920	0.1600	0.2380	0.0000
P	ChatGPT-detector-RoBERTa-Chinese	0.1511	0.1540	0.1920	0.1920	0.2320	0.1420	0.2180	0.1720
P	Detection-Longformer	0.0734	0.0740	0.0440	0.1540	0.0680	0.0680	0.1220	0.0820

Table 11: Per-test-language comparison of performance (TPR @ 5% FPR) of categories of MGT detectors (S – statistical, P – pretrained, F – finetuned). For readability, the finetuned category includes only the best performing combination of train languages of each base model. Bold represents the highest value per each test language.

Detector	Subset	All	cs	de	hr	hu	pl	sk	sl
Llama-3.2-3B	original	0.8821	0.9200	0.8850	0.9550	0.9500	0.9300	0.8350	0.8400
	paraphrased	**0.9464**	**0.9700**	**0.9400**	**0.9700**	**0.9700**	**0.9550**	**0.9400**	**0.9200**
	homoglyph	0.6357	0.8050	0.5350	0.8750	0.7500	0.7200	0.5850	0.5350
mDeBERTa-v3-base	original	0.8500	0.9100	0.8400	0.7600	0.9450	0.8300	0.8550	0.7000
	paraphrased	**0.9107**	**0.9600**	**0.9000**	**0.8700**	**0.9650**	**0.8850**	**0.9550**	**0.7800**
	homoglyph	0.1936	0.3800	0.1050	0.1950	0.2400	0.1400	0.2600	0.0300
Gemma-2-2B	original	0.8243	0.8950	0.8800	0.8600	0.9300	**0.9100**	0.8150	0.7300
	paraphrased	**0.9029**	**0.9600**	**0.9450**	**0.9500**	**0.9650**	**0.9100**	**0.9000**	**0.8450**
	homoglyph	0.5414	0.6700	0.6600	0.6200	0.6950	0.5650	0.4250	0.3400
XLM-RoBERTa-base	original	0.7529	0.8550	0.6750	0.8500	0.8800	0.7700	0.7800	0.6650
	paraphrased	**0.8100**	**0.8850**	**0.7100**	**0.8900**	**0.9350**	**0.7950**	**0.8650**	**0.7250**
	homoglyph	0.0821	0.1750	0.0250	0.1100	0.1650	0.0800	0.0650	0.0550
Fast-DetectGPT	original	**0.4400**	**0.4550**	**0.5350**	**0.4900**	**0.4850**	**0.4600**	**0.3450**	**0.4000**
	paraphrased	0.2464	0.2900	0.3600	0.2950	0.1300	0.1900	0.2650	0.2550
	homoglyph	0.0014	0.0000	0.0000	0.0050	0.0000	0.0050	0.0050	0.0050
Binoculars	original	**0.4307**	**0.5150**	**0.4950**	**0.5150**	**0.4900**	**0.5100**	**0.3650**	**0.3900**
	paraphrased	0.2600	0.3600	0.3600	0.3100	0.1600	0.2400	0.3150	0.2450
	homoglyph	0.0036	0.0050	0.0000	0.0050	0.0000	0.0100	0.0050	0.0050
LLM-Deviation	original	**0.2071**	**0.3550**	**0.2850**	**0.4400**	**0.2650**	**0.3700**	**0.2250**	**0.4050**
	paraphrased	0.0943	0.2450	0.1450	0.3600	0.0350	0.1600	0.2000	0.3050
	homoglyph	0.0036	0.0000	0.0000	0.0150	0.0000	0.0000	0.0100	0.0100
BLOOMz-3B-mixed-detector	original	0.1657	0.1850	0.1600	0.2250	0.2200	0.1300	0.2150	0.1250
	paraphrased	**0.5036**	**0.5250**	**0.4250**	**0.5550**	**0.5350**	**0.5150**	**0.5600**	**0.4100**
	homoglyph	0.0379	0.0500	0.0200	0.0500	0.0650	0.0500	0.0600	0.0050
ChatGPT-detector-RoBERTa-Chinese	original	0.1479	**0.1750**	**0.1850**	0.2050	0.1950	0.1850	0.2750	**0.1900**
	paraphrased	**0.1607**	0.1300	0.1300	**0.2250**	**0.2700**	**0.2000**	**0.2950**	0.1450
	homoglyph	0.0986	0.1200	0.0300	0.1600	0.1550	0.1250	0.2100	0.1400
Detection-Longformer	original	**0.0714**	**0.0550**	**0.0250**	**0.1350**	**0.0550**	**0.0800**	**0.0900**	**0.0600**
	paraphrased	0.0450	0.0500	0.0150	0.0900	0.0300	0.0750	0.0500	0.0450
	homoglyph	0.0179	0.0350	0.0000	0.0650	0.0000	0.0350	0.0150	0.0200

Table 12: Per-test-language comparison of performance (TPR @ 5% FPR) of the selected MGT detectors on original and adversarial data. Bold represents the highest value per each test language and each detector (in regard to original, paraphrased, and homoglyph subset).