mdok of KInIT: Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection

Notebook for the PAN Lab at CLEF 2025

Dominik Macko

Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia

Abstract

The large language models (LLMs) are able to generate high-quality texts in multiple languages. Such texts are often not recognizable by humans as generated, and therefore present a potential of LLMs for misuse (e.g., plagiarism, spams, disinformation spreading). An automated detection is able to assist humans to indicate the machine-generated texts; however, its robustness to out-of-distribution data is still challenging. This notebook describes our **mdok** approach in robust detection, based on fine-tuning smaller LLMs for text classification. It is applied to both subtasks of Voight-Kampff Generative AI Detection 2025, providing remarkable performance (**1st rank**) in both, the binary detection as well as the multiclass classification of various cases of human-AI collaboration.

Keywords

PAN 2025, Voight-Kampff Generative AI Detection 2025, Large language models, Machine-generated text detection, AI-content detection

1. Introduction

Continuously increasing quality of texts generated by artificial intelligence (AI) technology, such as large language models (LLMs), causes that humans are no longer able to differentiate between human-written and high-quality machine-generated texts. Naturally, this arises concerns about potential LLM misuse, e.g., for accelerated generation of disinformation [1, 2], plagiarism [3], or frauds for academic exams [4]. The automated means, also utilizing AI technology, are able to help humans to differentiate such texts; however, the automated detection performance is also not perfect (errors occur, such as false positives or false negatives). Further challenges are in application of the trained detectors to out-of-distribution data, i.e. data that are significantly different to data used for training (surprise data).

The *Voight-Kampff Generative AI Detection 2025* shared task [5], as a part of *PAN* lab [6] at the *CLEF 2025* conference, addresses two challenges in the machine-generated text detection problem area in two subtasks. Subtask 1 is focused on classical binary detection task distinguishing between machine-authored (class 1) and human-authored (class 0) texts, which addresses also the robustness of detectors to obfuscation and other surprise data. Subtask 2 is focused on human-AI collaborative text classification, covering six classes of collaboration, namely class 0 representing fully human-written texts; class 1 representing human-written, then machine-polished texts, class 2 representing machine-written, then machine-humanized texts, class 3 representing human-initiated, then machine-continued texts, class 4 representing deeply-mixed text (where some parts are written by a human and some are generated by a machine), and class 5 representing machine-written, then human-edited texts.

In this notebook, we describe our mdok approach addressing both subtasks (two separate systems). We build on our previously developed robust LLM fine-tuning for sequence classification [7], but adjust the training process to the shared task training data and binary/multiclass problem formulation. We further explore the usage of the most recent LLMs, size of which ranging from 1B to 14B parameters. The submitted systems are based on Qwen3-4B and Qwen3-14B models [8]. For the replication purpose, we publish the source code of the mdok approach[1].

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

dominik.macko@kinit.sk (D. Macko)

[1] https://github.com/kinit-sk/mdok

The contributions of the proposed approach are as follows:

- We have **proposed the usage of most modern Qwen3 LLM in robust mdok fine-tuning pipeline and evaluated** its robustness in sensitivity to obfuscation.
- We have **proposed a modification of robust fine-tuning mdok approach for multiclass** detection of AI-human collaboration.
- We have **benchmarked multiple most modern LLMs** on the provided evaluation datasets, making comparison of other models and approaches straightforward and fair.

2. Background

A binary machine-generated text detection is a well researched task, typically addressed by stylometric methods (e.g., a machine learning classifier trained on TF-IDF features), statistical methods (e.g., utilizing perplexity, entropy, or likelihood) [9, 10], or fine-tuned language models for classification task (e.g., by supervised or contrastive learning) [11, 12]. Most of the detection methods can be directly applied by existing frameworks, such as MGTBench [13] or IMGTB [14].

A multiclass machine-generated text classification is mostly researched in related authorship attribution task, identifying the author (generator) of the text. The problem of different classes of AI-human collaboration can be approached by existing authorship attribution methods [15, 16], only slightly different from binary (two-class) methods.

The robustness of machine-generated text detection methods against authorship obfuscation methods has been explored in [17]. Although it has been focused on multilingual settings and the results differ among languages, it has shown that fine-tuned detection methods are more robust against obfuscation than the statistical methods, while offering significantly higher detection performance. Further it has shown that including obfuscated data into fine-tuning process increases the detector's robustness against obfuscation. Such approach has been proposed in [7] and shown to increase even the generalization to out-of-distribution data. The training data mixture consists of social-media texts from the MultiSocial [18] dataset, news articles from the MULTITuDE [19] dataset, and obfuscated texts from [17]. For validation (i.e., model-checkpoint selection), the approach uses a unique massively multi-generator (75 generators) multilingual (7 languages) data of MIX2k composition of 18 existing labeled datasets to represent out-of-distribution data.

3. System Overview

For this shared task, we aimed to keep the system as simple as possible, ideally resulting in a single-model detection system for each task (avoiding ensembles), just by tweaking the data and the training process.

Although Subtask 1 does not explicitly mention any limitation in usage of additional training data, the Subtask 2 conditions are clear in this regard, not allowing additional training data. Therefore, for adoption of robust fine-tuning approach of [7], we do not include additional training data. Instead, for Subtask 1, we combine train and validation sets into training data and modify a small portion of them by using a generic homoglyph attack of [17]. For validation (model selection), we use MIX2k dataset only to ensure the best generalization to out-of-distribution data. For Subtask 2, we also combine train and validation sets for training, but we leave a pseudo-randomly balanced (500 samples per class) holdout portion for validation (in order to minimize bias due to imbalanced evaluation). An overview of this approach is illustrated in Figure 1.

For the fine-tuning process, we have used the adjusted published script for QLoRA based robust fine-tuning of [7]. Besides of modification of training data mixture, we have used learning rate of $2e-4$, avoided gradient accumulation, and limited the training time to 3 epochs. Due to already incorporated weighted cross entropy for loss calculations, it was straightforward to extend it to multiclass classification by calculating weights based on the training data class distribution.

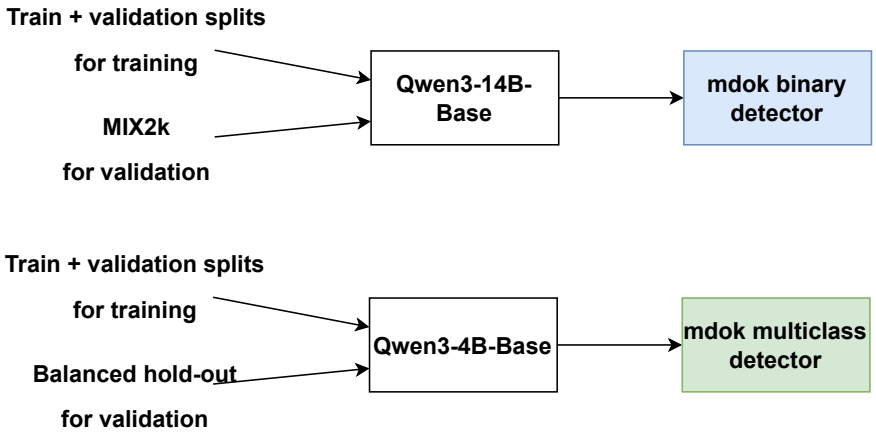

Figure 1: An overview of mdok approach for binary (top) and multiclass (bottom) detection of machine-generated text.

Since the evaluation on the original validation sets is not usable (due to data leakage) when combined in the training, we have used the original training sets for training to select suitable LLMs to robustly fine-tune for the two subtasks and compared them on validation sets (as well as on MIX2k data). Based on such comparison, we have decided to use the Qwen3-14B model for binary detection in Subtask 1 and the Qwen3-4B for multiclass classification in Subtask 2.

3.1. Homoglyph Attack

As mentioned, we have used a generic homoglyph attack of [17] to modify (in the pre-processing step) a small portion of training data (pseudo-randomly selected 10% of machine-generated texts) for Subtask 1 to increase its robustness to obfuscation. It uses the whole confusables table[2] to psudorandomly replace letters for their homoglyphs. The probability of a character to be replaced was set to 0.05 (i.e., about 5% of characters). It further integrates a pseudorandom insertion of visually unseenable zero-width-joiner character in the text (also with a probability of 0.05). We have used the random seed of 42. It has been previously shown that seeing such obfuscated data in the training (fine-tuning) effectively eliminates their highly negative effect on the detection.

3.2. MIX2k Dataset

For validation in out-of-distribution settings in Subtask 1, we have used MIX2k dataset, introduced by [7]. It contains 1,000 samples for each class (human and machine), pseudo-randomly sampled from 18 existing labeled datasets. Such data composition consists of data from 75 generators in 7 languages, thus being mostly out-of-distribution (in comparison to training data). Validation on such a dataset during fine-tuning procedure helps to select the model checkpoint, which has the best generalization to out-of-distribution data, effectively avoiding overfit to training data.

4. Results

The results are provided by using standard/official evaluation metrics specified for the shared task. Namely, in Subtask 1, the following metrics are used:

- **ROC-AUC**: The area under the ROC (Receiver Operating Characteristic) curve.
- **Brier**: The complement of the Brier score (mean squared loss).
- **C@1**: A modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases.

[2] https://www.unicode.org/Public/security/8.0.0/confusables.txt

Table 1
Comparison of Various Models and Variants on Validation Set of Subtask 1

Detector	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
TF-IDF baseline	0.996	0.951	0.984	0.980	0.981	0.978
PPMD baseline	0.786	0.799	0.757	0.812	0.778	0.786
Binoculars baseline	0.918	0.867	0.844	0.872	0.882	0.877
gemma-2-2b	0.9976	0.9980	0.9980	0.9985	0.9981	0.9981
gemma-2-9b-it	**0.9996**	**0.9994**	**0.9994**	**0.9996**	**0.9998**	**0.9996**
Qwen3-4B-Base	0.9991	0.9989	0.9989	0.9991	0.9997	0.9991
Qwen3-8B-Base	**0.9996**	**0.9994**	**0.9994**	**0.9996**	**0.9998**	**0.9996**
Qwen3-14B-Base	0.9994	0.9992	0.9992	0.9994	0.9997	0.9994
mdok (binary)	0.9972	0.9978	0.9978	0.9983	0.9978	0.9978

Table 2
Comparison of Various Models and Variants on MIX2k Out-of-distribution Dataset

Detector	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
TF-IDF baseline	0.5342	0.6817	0.5500	0.5837	0.5586	0.5817
PPMD baseline	0.4838	0.6970	0.5395	0.2614	0.4100	0.4783
Binoculars baseline	0.6368	**0.7443**	0.6440	0.5346	0.6554	0.6430
gemma-2-2b	0.6230	0.6230	0.6230	0.4512	0.6210	0.5882
gemma-2-9b-it	0.5920	0.5920	0.5920	0.3655	0.5480	0.5379
Qwen3-4B-Base	0.5710	0.5710	0.5710	0.3500	0.5066	0.5139
Qwen3-8B-Base	0.5720	0.5720	0.5720	0.2902	0.4797	0.4972
Qwen3-14B-Base	0.6480	0.6480	0.6480	0.5539	0.6597	0.6315
mdok (binary)	**0.6995**	0.6995	**0.6995**	**0.6696**	**0.7121**	**0.6960**

- **F1**: The harmonic mean of precision and recall.
- **F0.5u**: A modified F0.5 measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives.
- The arithmetic **mean** of all the metrics above.

For Subtask 2, standard metrics utilize **macro** averages of **Recall**, **Precision**, and **F1** scores, as well as overall **Accuracy**.

In Table 1, the detection performance for official validation set of Subtask 1 for each tested detector is provided. The results indicate that all the fine-tuned models perform almost ideally (very close to 1 values, outperforming all the baselines), which also indicates a potential overfit to in-distribution data.

Therefore, we have also tested the detectors on the MIX2k out-of-distribution dataset (see Table 2). As the results show, the Qwen3-14B-Base fine-tuned detector generalizes the best to out-of-distribution data. Therefore, it was our natural selection for robust fine-tuning using the mdok approach (described in Section 3), resulting in the mdok (binary) detector, which further boosted the performance, outperforming the baseline detectors.

For Subtask 2, the results for the official validation set are provided in Table 3. Although the results indicate that the Qwen2.5-1.5B models performs the best in official metric of Macro Recall, we selected newer and bigger Qwen3-4B-Base (with similar performance) for robust fine-tuning. Using the mdok approach for the fine-tuning resulted in mdok (multiclass) detector, hopefully outperforming the other variants. However, since it uses combined training and validation sets for training, the results indicated in Table 3 for this detector are not representative.

Table 3
Comparison of Various Models and Variants on Validation Set of Subtask 2

Detector	Macro Recall	Macro Precision	Macro F1	Accuracy
roberta-base baseline	0.7006	0.6515	0.6039	0.5593
Llama-3.2-1B	0.6172	0.6204	0.5558	0.5385
Qwen2.5-1.5B	0.7676	0.6466	0.6325	0.5848
Qwen3-1.7B-Base	0.6348	0.6763	0.5892	0.5466
Qwen2.5-3B	0.6911	0.6072	0.5768	0.6099
Qwen3-4B-Base	0.7345	0.6267	0.6311	0.6387
gemma-2-9b-it	0.6812	0.5634	0.5609	0.5826
mdok (multiclass)	**0.9727**	**0.9523**	**0.9617**	**0.9934**

Table 4
Official Leaderboard of Subtask 1 (for Teams Outperforming the Best-performing Baseline)

Rank	Team Name	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
1	**mdok (binary)**	0.853	**0.896**	**0.894**	**0.898**	**0.903**	**0.899**
2	steely	0.842	0.879	0.877	0.865	0.881	0.880
3	nexus-interrogators	**0.865**	0.874	0.870	0.860	0.881	0.879
4	yangjlg	0.845	0.878	0.871	0.856	0.881	0.877
5	cnlp-nits-pp	0.825	0.873	0.873	0.854	0.882	0.874
6	unibuc-nlp	0.828	0.885	0.864	0.845	0.876	0.872
7	moadmoad	0.822	0.866	0.865	0.855	0.882	0.871
8	iimasnlp	0.838	0.868	0.856	0.851	0.877	0.869
9	bohan-li	0.848	0.858	0.852	0.847	0.870	0.866
10	advacheck	0.802	0.855	0.855	0.854	0.879	0.863
11	hello-world	0.838	0.871	0.836	0.827	0.862	0.856

4.1. Subtask 1 Official Results

The official results are provided in Table 4, where Mean of the official metrics has been used for ranking. The results show that our approach outperforms all the others in all metrics except of ROC-AUC, where the system of one team outperformed our detector. The performances of the systems are close enough to yield a more efficient detection system, since we have relied on a rather large (14B of parameters) model, consuming more resources for inference. However, the mdok single-model detection system can still be more efficient than some more complex ensemble systems, relying on multiple inference rounds of multiple (although smaller) models. If the labeled test set will be released, an ablation study could pinpoint the strong parts of our system that make it the best performing and/or identify a smaller model to be used instead to provide a more efficient solution.

4.2. Subtask 2 Official Results

The official results are provided in Table 5, where Macro Recall is the official ranking metric. The results show that our approach outperforms all the others by a margin of 3%. Our decision to keep the system simple and focus on the training data and fine-tuning process paid off. If the labeled test set will be released, the further analysis and ablation study reveals whether we have made the right selection of the system to submit. However, even the top performance of our system (Macro Recall of 64.46%) is far from perfect (Macro Recall of 100%). This provides further space for improvements.

Table 5
Official Leaderboard of Subtask 2 (for Teams Outperforming the Baseline)

Rank	Team Name	Macro Recall	Macro F1	Accuracy
1	**mdok (multiclass)**	**64.46%**	**65.06%**	**74.09%**
2	lbh-1130	61.72%	61.73%	69.28%
3	anastasiya.vozniuk	60.16%	60.85%	69.04%
4	Gangandandan	57.46%	56.31%	66.81%
5	Atu	56.87%	56.45%	66.30%
6	TaoLi	56.74%	55.39%	66.27%
7	adugeen	56.11%	55.25%	64.79%
8	Real_Yuan	54.49%	54.40%	62.89%
9	WeiDongWu	54.09%	53.57%	63.01%
10	zhangzhiliang	54.06%	52.81%	61.65%
11	annepaka22	54.05%	53.49%	62.23%
12	a.dusuki	52.83%	51.44%	60.45%
13	padraig7	52.14%	51.81%	59.88%
14	a.elnenaey	49.56%	50.10%	58.96%
15	Maria Paz	49.19%	48.50%	56.94%
	roberta-base baseline	48.32%	47.82%	57.09%

5. Conclusion

The robustness of machine-generated text detection methods to out-of-distribution data is still challenging. However, we can cope with this problem by better mixture of training data, as we have show in the proposed mdok approach for both binary and multiclass detection. The resulting detectors are very competitive (**ranking 1st** in both subtasks), outperforming all the baselines, and promising better generalization to out-of-distribution data. Due to gold labels of the test data are not available yet, any ablation study pinpointing crucial parts of the systems and their specific effects was not possible. Further work is still needed to refine the process to achieve ideal performance.

Acknowledgments

This work was partially supported by the European Union NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00059 and partially by *LorAI – Low Resource Artificial Intelligence*, a project funded by Horizon Europe under GA No.101136646. We acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] I. Vykopal, M. Pikuliak, I. Srba, R. Moro, D. Macko, M. Bielikova, Disinformation capabilities of large language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 14830–14847. URL: https://aclanthology.org/2024.acl-long.793/. doi:10.18653/v1/2024.acl-long.793.

[2] A. Zugecova, D. Macko, I. Srba, R. Moro, J. Kopal, K. Marcincinova, M. Mesarcik, Evaluation of llm vulnerabilities to being misused for personalized disinformation generation, 2024. URL: https://arxiv.org/abs/2412.13666. arXiv:2412.13666.

[3] J. P. Wahle, T. Ruas, F. Kirstein, B. Gipp, How large language models are transforming machine-paraphrase plagiarism, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 952–963. URL: https://aclanthology.org/2022.emnlp-main.62/. doi:10.18653/v1/2022.emnlp-main.62.

[4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[5] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsivgun, J. Su, Z. Xie, M. Abassy, J. Mansurov, R. Xing, M. N. Ta, K. A. Elozeiri, T. Gu, R. V. Tomar, J. Geng, E. Artemova, A. Shelmanov, N. Habash, E. Stamatatos, I. Gurevych, P. Nakov, M. Potthast, B. Stein, Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[6] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[7] D. Macko, R. Moro, I. Srba, Increasing the robustness of the fine-tuned multilingual machine-generated text detectors, 2025. URL: https://arxiv.org/abs/2503.15128. arXiv:2503.15128.

[8] Q. Team, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[9] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting llms with binoculars: Zero-shot detection of machine-generated text, 2024. URL: https://arxiv.org/abs/2401.12070. arXiv:2401.12070.

[10] G. Bao, Y. Zhao, Z. Teng, L. Yang, Y. Zhang, Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature, in: The Twelfth International Conference on Learning Representations, 2023.

[11] M. Spiegel, D. Macko, KInIT at SemEval-2024 task 8: Fine-tuned LLMs for multilingual machine-generated text detection, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 558–564. URL: https://aclanthology.org/2024.semeval-1.84/. doi:10.18653/v1/2024.semeval-1.84.

[12] S. R. Dipta, S. Shahriar, Hu at semeval-2024 task 8a: Can contrastive learning learn embeddings to detect machine-generated text?, 2024. URL: https://arxiv.org/abs/2402.11815. arXiv:2402.11815.

[13] X. He, X. Shen, Z. Chen, M. Backes, Y. Zhang, Mgtbench: Benchmarking machine-generated text detection, in: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 2251–2265. URL: https://doi.org/10.1145/3658644.3670344. doi:10.1145/3658644.3670344.

[14] M. Spiegel, D. Macko, IMGTB: A framework for machine-generated text detection benchmarking, in: Y. Cao, Y. Feng, D. Xiong (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 172–179. URL: https://aclanthology.org/2024.acl-demos.17/. doi:10.18653/v1/2024.acl-demos.17.

[15] A. Uchendu, Z. Ma, T. Le, R. Zhang, D. Lee, TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2001–2016. URL: https:

//aclanthology.org/2021.findings-emnlp.172/. doi:10.18653/v1/2021.findings-emnlp.172.

[16] L. La Cava, D. Costa, A. Tagarelli, Is contrasting all you need? contrastive learning for the detection and attribution of ai-generated text, in: ECAI 2024, IOS Press, 2024, pp. 3179–3186.

[17] D. Macko, R. Moro, A. Uchendu, I. Srba, J. S. Lucas, M. Yamashita, N. I. Tripto, D. Lee, J. Simko, M. Bielikova, Authorship obfuscation in multilingual machine-generated text detection, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 6348–6368. URL: https://aclanthology.org/2024.findings-emnlp.369/. doi:10.18653/v1/2024.findings-emnlp.369.

[18] D. Macko, J. Kopal, R. Moro, I. Srba, MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts, 2024. URL: https://arxiv.org/abs/2406.12549. arXiv:2406.12549.

[19] D. Macko, R. Moro, A. Uchendu, J. Lucas, M. Yamashita, M. Pikuliak, I. Srba, T. Le, D. Lee, J. Simko, M. Bielikova, MULTITuDE: Large-scale multilingual machine-generated text detection benchmark, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 9960–9987. URL: https://aclanthology.org/2023.emnlp-main.616/. doi:10.18653/v1/2023.emnlp-main.616.