AI text detection for under-resourced languages*

Jernej Vičič[1,*,†]

[1]Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška 8, 6000 Koper, Slovenia

Abstract

The rapid development of large language models (LLMs) has created a growing need for methods to distinguish AI-generated text from human-written text. Most existing detection methods are developed and evaluated on English, and authors report performance decreases for other languages, particularly morphologically rich and under-represented languages. This study focuses on three Slavic languages: Czech, Slovak, and Slovene, and introduces a controlled news-domain dataset containing human-written articles and corresponding texts generated by multiple LLMs. As a reference experiment, a TF–IDF and Multinomial Naive Bayes detector is evaluated on English and Slavic datasets, demonstrating substantial variation across data distributions and text generators. The study further investigates a multilingual adaptation of BENADV, a detection approach based on conformity to Benford's law. Experiments are designed to determine whether Benford-based features distinguish human-written from machine-generated text across languages, whether they support cross-lingual transfer, and how multilingual and language-specific embedding models affect detection performance. The study provides a dataset and experimental framework for investigating AI-text detection in under-represented Slavic languages.

Keywords

AI-generated text, AI-text detection, Benford's law, multilingual NLP, low-resource languages

1. Introduction

Recent advances in large language models (LLMs) have raised growing concerns about the reliable identification of machine-generated text. A large number of detection methods have been proposed; their development and evaluation have largely centered on English, and their effectiveness across other languages is comparatively limited. The presented research focuses on three Slavic languages: Czech, Slovak, and Slovene, which share many linguistic characteristics but also exhibit important language-specific properties. Mainly, there is a noticeable gap in available datasets and tools between Czech and the other two languages. A need for a new dataset of parallel human text and AI-generated text for all three languages was clear after a search of the scientific literature. The first version of the dataset is the first outcome of the presented research. We also investigate a simple but not computationally cheap and interpretable detection approach based on Benford's law and transformer embeddings. We hypothesize that AI-generated texts follow Benford's law more closely than human-written texts when numerical patterns are extracted from contextual embedding representations. We also hypothesize that the investigated method transfers well across languages and domains. To test this hypothesis, token-level embeddings are obtained from multilingual and language-specific transformer models and transformed into leading-digit distributions. The similarity between the observed distributions and Benford's law is measured using Kullback–Leibler divergence, chi-square statistics, and mean squared error. The proposed framework is evaluated on multilingual datasets, purposefully tailored for the ITAT conference, containing English, Slovene, Czech, and Slovak texts. The experiments examine whether the Benford-law hypothesis holds across languages, whether the approach can transfer between languages, and how different embedding models affect performance. Both multilingual and language-specific

ITAT 2026: Information Technologies – Applications and Theory, 2026

*Editors of the proceedings (editors): Lucie Ciencialová, Martin Holeňa, Róbert Jajcay, Tatiana Jajcayová, František Mráz, Richard Ostertág, Dana Pardubská, Martin Plátek, Martin Stanek

*Corresponding author.

jernej.vicic@upr.si (J. Vičič)

https://osebje.famnit.upr.si/~jernej// (J. Vičič)

0000-0002-7876-5009 (J. Vičič)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-09

CEUR-WS.org/Vol-4323/invited2.pdf

transformer models are evaluated. A simple TF-IDF representation with a Multinomial Naive Bayes classifier is used as a reference baseline along with results reported in previous studies. The goal of this study is to determine whether Benford-law-based features capture general characteristics of AI-generated text that are less dependent on language than existing approaches. The study investigates whether these statistical properties can provide a more language-independent basis for multilingual AI-text detection, particularly for languages that are under-represented in current research.

2. State of the Art

AI-text detection is feasible in controlled settings, but according to recently published results, post-hoc detectors are fragile when applied beyond the exact data they were built on.

One of the most cited benchmarks, MGTBench, states that metric-based and model-based detectors transfer better across LLMs and datasets, but that paraphrasing, random spacing, and adversarial perturbations can significantly diminish detection effectiveness [1]. A separate 2026 benchmark reports the same broad problem: transformer detectors can look near-perfect in distribution, yet they degrade under domain shift, and no method generalizes robustly across domains and LLM sources [2]. Binoculars [3] reports state-of-the-art zero-shot detection and says it can detect over 90% of generated samples from ChatGPT and other LLMs at a false positive rate of 0.01%. AuthentiGPT reports a 0.918 AUROC on a domain-specific dataset, [4] and T5LLMCipher claims an average 19.6% F1 gain on unseen generators and domains plus 93.6% attribution accuracy [5]. Those are encouraging, but they are still benchmark results, not proof that a detector will stay reliable on new models, lightly edited text, or mixed human-AI writing.

A social-media benchmark shows that humans can score highly on some curated data, with annotators distinguishing AI-generated from human text at 96.5% accuracy on one Zhihu dataset, while the authors note that real-platform detection is harder than simulated detection [6]. On the other side of the field, watermarking papers treat prevention more than forensics: they explicitly aim for watermark methods that remain robust against text editing and paraphrasing [7]. These approaches represent complementary strategies: post-hoc detection attempts to infer authorship from properties of the observed text, whereas watermarking introduces provenance signals during text generation.

2.1. State of the Art for non-English Languages

AI-text detectors can be adapted to non-English languages, but performance drops and error patterns shift once leaving the English language. Subramaniam [8] evaluates OpenAI's AI Text Classifier on English, German, Arabic, Hindi, Chinese, and Swahili texts and finds language-dependent errors: false-positive classifications were more prevalent for Hindi and Arabic, whereas texts in several other languages were frequently not classified with sufficient confidence. These concerns are further supported by the large-scale MULTITuDE benchmark by Macko et al. [9].

More recent work has increasingly focused on multilingual transformer models and task-specific fine-tuning for machine-generated text detection. Sharimbayev and Duisen [10] trained multilingual transformer models on a dataset of 98,224 samples across nine languages. Their best threshold-optimized XLM-RoBERTa model achieved 95.56% accuracy and an F1-score of 96.29%. Despite the strong overall multilingual performance, English exhibited the lowest language-specific accuracy, which the authors partly attribute to greater diversity in English writing styles. In another multilingual evaluation, Schaefer and Steinebach [11] investigated transformer fine-tuning, adapters, and adapter fusion. Their experiments showed that adapter fusion degraded performance in the multilingual setting, whereas a directly fine-tuned XLM-RoBERTa model achieved a macro F1-score of 0.7258 on the final multilingual test set.

Recent multilingual studies demonstrate that detector performance depends strongly on the detection approach. Quaremba et al. [12] introduced WETBench, a multilingual, multi-generator benchmark designed around realistic Wikipedia editing tasks. It achieved an average accuracy of approximately 78%, whereas zero-shot detectors achieved only 58%. The results also varied across languages, generation

models, and editing tasks, highlighting the limited robustness and generalizability of current machine-generated text detection methods.

Taken together, these results do not suggest that non-English detection is infeasible; rather, they indicate that multilingual generalization remains fragile, particularly for low-resource languages, code-mixed settings, paraphrased text, and domain shifts.

2.2. AI Text Detection for Czech and Slovak

The practical implications are particularly relevant for Central European languages. Šigut and Foltýnek [13] evaluated publicly available AI-text detection tools on Czech, Slovak, and English texts and found that only one of the evaluated tools, Compilatio, showed some ability to detect AI-generated Czech and Slovak texts. They additionally investigated whether translating Czech and Slovak texts into English prior to detection could compensate for the predominantly English-oriented nature of existing detectors. Their results demonstrate that detection behaviour is strongly affected by language and translation, making translation to English an unreliable general substitute for native-language detection. The study also found that texts generated by ChatGPT-4 were more difficult to detect than those generated by ChatGPT-3.5 [13].

More recent evidence supports this broader conclusion. The CEAID benchmark introduced by Macko and Kopal [14] systematically evaluates machine-generated text detection for Central European languages. The authors observe that existing detection methods are developed for English and can therefore be almost unusable for non-English languages when they rely solely on cross-lingual transfer. Their experiments further show that supervised detectors fine-tuned specifically on Central European languages achieve the strongest performance and are also the most robust against text obfuscation [14].

3. Methodology

To evaluate the cross-lingual robustness of Benford-law-based machine-generated text detection, we design a multilingual experimental framework targeting three Slavic languages (Slovene, Czech, and Slovak) and English as a reference. The experiments are structured progressively, beginning with the validation of the core statistical hypothesis and subsequently extending toward supervised and cross-lingual detection settings.

3.1. Datasets

Two English datasets, a relatively old (two years) but large dataset, *Comprehensive AI vs Human*, and a newer dataset from 2025, *Student or AI* [1] were used mostly for their simple availability on the Kaggle platform [2] as they were mostly used as reference. A search of the available datasets with parallel human and AI-generated text for the three targeted Slavic languages revealed no useful candidates.

Although there is a good resource for the Czech language, the AI Koditex [15], we did not identify comparable resources for Slovak and Slovene, although there were some attempts to create a multilingual comprehensive dataset for human-to-AI-generated data, such as MultiSocial [16]. Preliminary inspection indicated substantial quality limitations for the Slovene and Slovak subsets, making them unsuitable for the controlled experiments considered here. The Slavic portion of the languages was targeted by a purposefully composed dataset using Slovene, Czech, and Slovak texts described in Section 3.1.1.

3.1.1. AI-generated text detection for Slavic languages

We constructed a multilingual benchmark for the detection of AI-generated text in three Slavic languages: Czech, Slovak, and Slovene. The benchmark is initially restricted to the news domain, allowing tighter control over genre, register, document structure, and the information available to human and

[1] Adapted from: Syed Ali Haider: https://linktr.ee/Syed_Ali_Haider
[2] Kaggle: https://kaggle.com

machine authors. Rather than treating the resulting resource as a general-purpose AI-text benchmark, the current design targets the more precisely defined task of distinguishing human-written from machine-generated journalistic text. To reduce differences caused by fundamentally different types of publishers, public-service news organizations were selected as the initial human-text sources for each language. The resulting sources are RTVSLO (https://www.rtvslo.si/) for Slovene, ČT24 (Česká televize) (https://ct24.ceskatelevize.cz/) for Czech, and Správy STVR (https://spravy.stvr.sk/) for Slovak. This choice provides three broadly comparable professional journalistic environments and avoids constructing the human class from heterogeneous sources such as news articles, forums, essays, and social media. The decision to focus initially on a single domain was deliberate. A multi-domain corpus would increase coverage but would also introduce additional confounding variables, including differences in text length, register, editing practices, document structure, authorship provenance, and the way equivalent machine-generated texts can be produced. News articles provide a comparatively well-defined generation process in which an event or summary can serve as the information source for both human and synthetic articles.

This benchmark should be regarded as a news-domain benchmark, and performance obtained on it should not be interpreted as evidence of equivalent performance on unrestricted AI-generated text.

For each source article, two synthetic versions were generated using DeepSeek (DeepSeek-V4-Flash) and ChatGPT (GPT-5.6 Sol). Both generators received the same lead-conditioned generation instruction. Additionally, for Slovene, the GAMS (GaMS3-12B-Instruct) LLM was used.

The generation procedure uses the summary or lead associated with a human-written news article as input to a large language model (LLM). The LLM is then instructed to generate a complete news article based on this information. To reduce document length as a potential detection factor, the target length of the generated article is adapted to the length of the corresponding human-written article. The generated text is therefore constrained to approximately match the length of its source article rather than being produced within a single fixed length range for the whole dataset. Conceptually, each source item therefore produces a paired observation: `(lead, human article, LLM-generated article)`. Figure 1 shows an example template for the Slovene prompt for the generation of articles.

> This is the lead for a news article in the Slovenian language: lead, followed by the length in words of the original article: word_count. Please prepare the whole article using this information. Do not use the internet when preparing this answer. The length of the article should be approximately the same as the original article. The format of the article should be: Title, subtitle, lead, and the body of the article. The language of the article should be Slovene. Output the result as JSON: ”title”, ”subtitle”, ”lead”, ”body”.

Figure 1: An example template for the prompt creation – the Slovene part of the dataset. The target word count is dynamically set approximately to the length of the original article.

This paired construction has an important advantage over independently sampling human and synthetic documents: the two classes are conditioned on approximately the same topic and underlying event. Consequently, topic differences should be substantially reduced as a potential shortcut for the detector.

Several limitations of this procedure were identified during benchmark design. In particular, the lead may contain wording directly derived from the human article and therefore transfers some human lexical choices into the synthetic document. More importantly, the original journalist generally had access to substantially more information than is contained in the lead, including quotations, names, numerical information, background material, and contextual details. An LLM conditioned only on the lead may consequently produce generic elaboration or hallucinated details. A classifier could then distinguish the classes through differences in information density, quotations, named entities, or factual specificity rather than through properties of human and machine authorship itself.

Preliminary experiments additionally showed that document length could act as a strong shortcut for classification. In particular, generated texts from some models exhibited a substantially different

Detector	Dataset	Class	Prec./SD	Rec./SD	F1/SD	Supp.
Only length	SLO ChatGPT – no length control	0	0.73/0.06	0.84/0.06	0.78/0.05	100
		1	0.81/0.06	0.68/0.09	0.74/0.07	100
		Acc./SD	0.76/0.06			200
	SLK ChatGPT – no length control	0	0.77/0.06	0.87/0.08	0.81/0.04	98
		1	0.85/0.07	0.73/0.10	0.78/0.06	98
		Acc./SD	0.80/0.04			196
	CES ChatGPT – no length control	0	0.86/0.07	0.93/0.04	0.89/0.04	100
		1	0.92/0.04	0.84/0.09	0.88/0.06	100
		Acc./SD	0.89/0.05			200

Table 1
Results of a preliminary experiment with a simple detector based solely on article length, evaluated on the first version of the dataset without strict length control.

length distribution from the corresponding human-written articles. A control experiment using document length as the only classification feature confirmed that length alone could provide substantial discriminatory power. The values are presented in Table 1. To reduce this source of bias, the generation procedure was revised so that the target length of each generated article approximately follows the length of the original article.

For this reason, lead-conditioned article reconstruction is treated as a specific generation condition, rather than as a complete operational definition of AI-generated news. Additional generation conditions under consideration include reconstruction from structured factual information and rewriting of the original article.

Dedicated extraction pipelines were developed for all three news sources. Because each publisher uses a different HTML structure, source-specific extraction was preferred over generic webpage-to-text extraction. This provides substantially tighter control over which elements become part of the benchmark.

The lead and article body are deliberately retained as separate fields. This is essential for the planned generation methodology: the lead can be supplied to the LLM as conditioning information, while the body constitutes the corresponding human-written target text.

The source article together with the AI-generated versions must constitute the splitting unit, ensuring that the human article and all synthetic reconstructions of the same event remain in the same partition.

The dataset is available at the Zenodo repository: https://doi.org/10.5281/zenodo.21992521 [17].

3.2. Detection methods

3.2.1. Reference method

For the reference metric, we took an off-the-shelf simple detector based on TF-IDF [18] and Multinomial Naive Bayes [19] from an easily accessible Kaggle implementation [3]. Figure 2 shows the pipeline used for the reference metric implementation.

3.2.2. Benford's Law-based Method

BENADV, proposed by Wang et al. [20], is a machine-generated text detection method that exploits differences in the numerical distributions of contextual text representations. The main idea is based on Benford's law [21], hence the name. Instead of relying directly on lexical or stylistic properties, the method transforms a text into numerical representations obtained from pretrained language models (embeddings).

For a text consisting of n tokens and a hidden representation of dimension d, this produces an embedding matrix

[3]Adapted from: Syed Ali Haider: https://linktr.ee/Syed_Ali_Haider

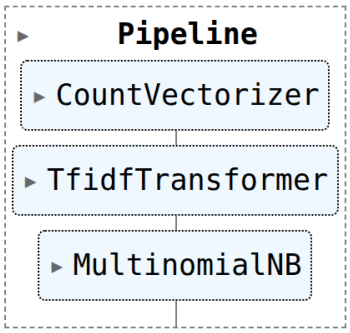


Figure 2: Reference binary classifier based on TF-IDF and Multinomial Naive Bayes (Kaggle implementation)

$$E(x) \in \mathbb{R}^{n \times d}. \tag{1}$$

The numerical values of this matrix are treated as observations from which the first significant digits are extracted. For every non-zero embedding value, its first significant digit $k \in \{1, \dots, 9\}$ is extracted. The empirical distribution of these digits is compared with the distribution predicted by **Benford's law**,

$$P(k) = \log_{10}\left(1 + \frac{1}{k}\right), \qquad k = 1, \dots, 9. \tag{2}$$

The most important observation reported by Wang et al. [20] is that machine-generated text (MGT) generally exhibits stronger conformity to Benford's law than human-written text (HWT). BENADV characterizes each text according to how closely its leading-digit distribution conforms to Benford's law. Several goodness-of-fit measures are used to quantify this conformity, including **Kullback–Leibler (KL) divergence** [22], **chi-squared distance** [23], and **mean squared error (MSE)** [24].

The original BENADV method additionally incorporates adversarial perturbations and a learned classification mechanism. These components are not included in the present experiments, which focus specifically on the Benford-conformity representation.

Consequently, the main BENADV processing pipeline can be summarized as

$$\begin{aligned} \text{text} &\rightarrow \text{language-model representations} \\ &\rightarrow \text{leading-digit distribution} \\ &\rightarrow \text{Benford conformity measures} \\ &\rightarrow \text{human/AI classification.} \end{aligned} \tag{3}$$

In our work, this methodology provides the basis for investigating whether the reported Benford-related differences are **language-independent**, by extending the analysis from English to Slovene, Czech, and Slovak.

3.3. Experiments

3.3.1. Experiment 1: Validation of the Benford Hypothesis

The first experiment investigates the central hypothesis underlying BENADV:

Machine-generated text conforms more closely to Benford's law than human-written text.

For each language independently, token-level contextual embeddings are extracted from multilingual transformer models. The embedding matrices are flattened into numerical vectors, from which the leading significant digits are obtained. The empirical digit distributions are then compared against the theoretical Benford distribution:

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right), \quad d \in \{1, \ldots, 9\}.$$

For every text sample, the following conformity metrics are computed:

- Kullback–Leibler divergence (KL),
- Chi-square statistic (χ^2),
- Mean squared error (MSE),

The primary objective is to determine whether AI-generated texts consistently exhibit stronger conformity to Benford's law than human-written texts across different languages.

3.3.2. Experiment 2: Cross-Language Transfer

The second experiment evaluates whether Benford-based representations generalize across languages.

A classifier is trained exclusively on English texts and subsequently evaluated on Slovene, Czech, and Slovak datasets. This setup tests whether the Benford-related statistical patterns represent universal characteristics of machine-generated text or whether they remain language-specific phenomena.

The experiment particularly investigates the impact of linguistic morphology and tokenizer fragmentation on transferability.

3.3.3. Experiment 3: Embedding Model Comparison

The third experiment compares the influence of different embedding models on Benford-law conformity and downstream detection performance. Planned language-specific models include GaMS for Slovene and Czert or RobeCzech for Czech.

The following categories of models are observed:

- Multilingual transformer models:
  - `xlm-roberta-base`,
  - `bert-base-multilingual-cased`,
  - `mT5`.
- Language-specific models:
  - GaMS (Slovene),
  - Czech-specific transformer models (e.g., Czert or RobeCzech),

The experiment aims to determine whether language-specialized embedding spaces produce stronger Benford-law separation than general multilingual representations.

3.3.4. Research Questions

The experiments are designed to answer the following research questions:

1. Does machine-generated text exhibit stronger conformity to Benford's law across multiple languages?
2. How strongly does language morphology influence Benford-based detection?
3. Can Benford-based detectors generalize across languages?
4. Do language-specific embedding models outperform multilingual models?

4. Results

For the reference metric, we took an off-the-shelf simple detector based on TF–IDF and Multinomial Naive Bayes as presented in Section 3.2.1. Table 2 shows the method's performance on in-domain detection across reference datasets and a newly prepared dataset. The datasets are presented in Section 3.1.

The reference classifier was evaluated using repeated stratified cross-validation. We used 5-fold stratified cross-validation repeated twice, resulting in ten train–test evaluations for each dataset. Stratification was used to preserve approximately the same class distribution in each fold. Stratification was performed without dividing the human/AI news pair into test/train sets. A fixed random seed was used to ensure reproducibility.

The complete text-processing and classification pipeline was fitted independently within each training fold. The pipeline first transformed the documents into bag-of-words representations using CountVectorizer, followed by TF–IDF weighting and classification with Multinomial Naive Bayes. Fitting the vectorizer and TF–IDF transformation exclusively on each training fold prevents information from the corresponding test fold from influencing feature construction.

Precision, recall, and F1-score were calculated separately for the human- and AI-generated classes in each test fold. Table 2 reports the mean and standard deviation (SD) of each metric across the ten evaluations. Accuracy was calculated independently for every test fold and is likewise reported as mean/SD across the ten evaluations. The support column reports the total number of samples of each class in the complete dataset rather than the accumulated test-fold support.

For the newly constructed Slavic datasets, human-written and AI-generated articles originating from the same source item were treated as related observations during dataset construction. The evaluation therefore focuses on the reference classifier's ability to distinguish human-written from machine-generated articles under the corresponding language and generator conditions.

The results show strong performance on both English reference datasets. On the Slavic datasets, performance depends strongly on the text generator. Accuracy for DeepSeek-generated texts is only 0.64–0.65 across Slovene, Slovak, and Czech, whereas it reaches 0.94–0.99 for ChatGPT-generated texts. Particularly notable is the low recall for human-written texts in the DeepSeek experiments (0.30–0.37), indicating that many human articles are incorrectly classified as AI-generated.

5. Discussion and Conclusion

This paper presents preliminary work toward evaluating machine-generated text detection in underrepresented Slavic languages. In addition to reviewing recent multilingual detection approaches, we introduce a controlled news-domain dataset containing human-written and LLM-generated articles in Slovene, Czech, and Slovak. The paired construction, in which synthetic articles are generated from the lead of the corresponding human-written article, reduces topic differences between the two classes while preserving a realistic news-generation setting.

The preliminary reference experiments demonstrate that detector performance can depend strongly on the text generator. The TF–IDF and Multinomial Naive Bayes classifier achieves high accuracy on the two English reference datasets. On the Slovene dataset, however, accuracy reaches only 0.64 when distinguishing human-written articles from DeepSeek-generated articles, compared with 0.97 for ChatGPT-generated articles. Particularly notable is the low recall of 0.30 for human-written texts in the DeepSeek experiment, while recall for AI-generated texts reaches 0.99.

These results indicate that high performance on established benchmarks does not necessarily imply robust performance under changes in language, dataset, or text generator.

The current results should be regarded as preliminary and do not yet establish whether Benford-law-based representations provide a more language-independent detection signal. They do, however, motivate the planned evaluation of the BENADV-inspired approach. Future experiments will examine whether differences in Benford conformity between human-written and machine-generated text persist

Detector	Dataset	Class	Precision/SD	Recall/SD	F1-score/SD	support
Reference	ENG AI vs Human	0	0.94/0.11	0.99/0.05	0.96/0.12	7295
		1	0.98/0.08	0.89/0.09	0.94/0.13	43886
		Accuracy/SD	0.95/0.10			51181
	ENG Student vs AI	0	0.99/0.02	0.99/0.01	0.99/0.02	550
		1	0.99/0.01	1.00/0.00	1.00/0.00	552
		Accuracy/SD	1.00/0.00			1102
	SLO DeepSeek	0	0.97/0.05	0.30/0.12	0.44/0.15	96
		1	0.59/0.04	0.99/0.02	0.74/0.03	96
		Accuracy/SD	0.64/0.06			192
	SLO ChatGPT	0	0.99/0.02	0.94/0.06	0.97/0.03	100
		1	0.95/0.05	0.99/0.02	0.97/0.02	100
		Accuracy/SD	0.97/0.02			200
	SLK DeepSeek	0	0.88/0.13	0.37/0.11	0.51/0.10	97
		1	0.60/0.05	0.93/0.07	0.73/0.04	97
		Accuracy/SD	0.65/0.06			194
	SLK ChatGPT	0	1.00/0.00	0.98/0.03	0.99/0.01	100
		1	0.98/0.02	1.00/0.00	0.99/0.01	100
		Accuracy/SD	0.99/0.01			200
	CES DeepSeek	0	0.96/0.09	0.30/0.09	0.45/0.12	98
		1	0.58/0.04	0.98/0.03	0.73/0.03	98
		Accuracy/SD	0.64/0.05			196
	CES ChatGPT	0	0.97/0.03	0.91/0.07	0.94/0.03	101
		1	0.92/0.05	0.97/0.03	0.94/0.02	101
		Accuracy/SD	0.94/0.03			202

Table 2
Evaluation results of the Kaggle reference implementation on all the presented datasets.

across Slovene, Czech, Slovak, and English, and whether these properties transfer across languages and embedding models. The dataset and experimental framework presented here provide a basis for this continuing investigation of multilingual AI-text detection.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] X. He, X. Shen, Z. Chen, M. Backes, Y. Zhang, Mgtbench: Benchmarking machine-generated text detection, 2023. URL: https://dl.acm.org/doi/pdf/10.1145/3658644.3670344. doi:10.1145/3658644.3670344.

[2] M. S. Baidya, S. Baidya, C. Chawla, Detecting the machine: A comprehensive benchmark of ai-generated text detectors across architectures, domains, and adversarial conditions, 2026. URL: https://www.semanticscholar.org/paper/80093f769169cf5618110132af39242ecbc522f3.

[3] 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://www.semanticscholar.org/paper/8dafb105a0cec1956443003eeea9fcf13c0ffa99.

[4] Z. Guo, S. Yu, Authentigpt: Detecting machine-generated text via black-box language models denoising, 2023. URL: https://arxiv.org/pdf/2311.07700. doi:10.48550/arXiv.2311.07700.

[5] M. Bethany, B. Wherry, E. Bethany, N. Vishwamitra, P. Najafirad, Deciphering textual authenticity: A generalized strategy through the lens of large language semantics for detecting human vs. machine-generated text, 2024. URL: https://arxiv.org/pdf/2401.09407. doi:10.48550/arXiv.2401.09407.

[6] W. Cui, L. Zhang, Q. Wang, S. Cai, Who said that? benchmarking social media ai detection, 2023. URL: https://arxiv.org/pdf/2310.08240. doi:10.48550/arXiv.2310.08240.

[7] X. Zhao, P. Ananth, L. Li, Y.-X. Wang, Provable robust watermarking for ai-generated text, 2023. URL: https://arxiv.org/pdf/2306.17439. doi:10.48550/arXiv.2306.17439.

[8] R. Subramaniam, Identifying text classification failures in multilingual ai-generated content, International Journal of Artificial Intelligence & Applications 14 (2023) 57–63. doi:10.5121/ijaia.2023.14505.

[9] D. Macko, R. Moro, A. Uchendu, J. S. 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: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, pp. 9960–9987. doi:10.18653/v1/2023.emnlp-main.616.

[10] B. Sharimbayev, S. Duisen, Cross-lingual detection of machine-generated text using multilingual transformers and balanced data, in: 2026 18th International Conference on Electronics, Computer and Computation (ICECCO), 2026. doi:10.1109/ICECCO67619.2026.11488773.

[11] K. Schaefer, M. Steinebach, Fraunhofer SIT at GenAI detection task 1: Adapter fusion for AI-generated text detection, in: Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), Int Conf on Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 178–183.

[12] G. Quaremba, E. Black, D. Vrandečić, E. Simperl, WETBench: A benchmark for detecting task-specific machine-generated text on wikipedia, in: Proceedings of the 2nd Workshop on Advancing Natural Language Processing for Wikipedia (WikiNLP 2025), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 10–30. doi:10.18653/v1/2025.wikinlp-1.6.

[13] P. Šigut, T. Foltýnek, Can we detect ChatGPT-generated texts in czech and slovak languages?, in: Proceedings of the Sixteenth Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2023), 2023.

[14] D. Macko, J. Kopál, CEAID: Benchmark of multilingual machine-generated text detection methods for central european languages, in: Findings of the Association for Computational Linguistics: ACL 2026, Association for Computational Linguistics, 2026. URL: https://aclanthology.org/2026.findings-acl.1458/. doi:10.18653/v1/2026.findings-acl.1458.

[15] J. Milička, A. Marklová, V. Cvrček, Ai brown and ai koditex: Llm-generated corpora comparable to traditional corpora of english and czech texts, arXiv preprint arXiv:2509.22996 (2025).

[16] D. Macko, J. Kopal, R. Moro, I. Srba, 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), 2025, pp. 727–752.

[17] J. Vičič, D. Vake, Ai-generated text detection for slavic languages, 2026. URL: https://doi.org/10.5281/zenodo.21992521. doi:10.5281/zenodo.21992521.

[18] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (1988) 513–523. doi:10.1016/0306-4573(88)90021-0.

[19] A. McCallum, K. Nigam, A comparison of event models for naive bayes text classification, in: AAAI-98 Workshop on Learning for Text Categorization, 1998, pp. 41–48.

[20] Z. Wang, C. Zhang, M. Ren, A novel benford's law-driven approach for detecting machine-generated text, ACM Transactions on Information Systems 43 (2025) 1–25. doi:10.1145/3748305.

[21] F. Benford, The law of anomalous numbers, Proceedings of the American Philosophical Society 78 (1938) 551–572.

[22] S. Kullback, R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics 22 (1951) 79–86. doi:10.1214/aoms/1177729694.

[23] K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50 (1900) 157–175. doi:10.1080/14786440009463897.

[24] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2 ed., Springer, New York, 2009. doi:10.1007/978-0-387-84858-7.