BuST: A Siamese Transformer Model for AI Text Detection in Bulgarian

Andrii Maslo **Silvia Gargova**
Big Data for Smart Society Institute (GATE), Bulgaria,
andri.maslo@gate-ai.eu, silvia.gargova@gate-ai.eu

Abstract

The rapid advancement of large language models (LLMs) has made machine-generated text increasingly indistinguishable from human-written content, posing significant challenges for reliable detection. In this study, we propose BuST (Bulgarian Siamese Transformer), a novel detection methodology tailored for Bulgarian-language text that leverages paraphrase-based semantic similarity to identify machine-generated content. Inspired by the RAIDAR approach, BuST utilizes a Siamese Transformer architecture to compare original texts with their LLM-generated paraphrases, capturing subtle linguistic divergences indicative of synthetic origin. Our pilot experiments demonstrate that BuST effectively learns fine-grained patterns of semantic (mis)alignment, achieving an accuracy of 88.79% and an F1-score of 88.0%, reflecting competitive performance relative to strong baselines. While a pretrained BERT model achieved the highest overall accuracy (93.7%) and F1 score (93.9%), BuST's paraphrase similarity learning provides a promising, model-agnostic framework adaptable to under-resourced languages. These results highlight the potential of paraphrase-based methods as a robust strategy for machine-generated text detection.

1 Introduction

The rapid advancement of large language models (LLMs) has enabled them to generate text that closely resembles human writing. While this progress has fueled applications in education, customer support, and creative industries, it also raises serious risks: misinformation, fake reviews, and impersonation can be easily produced and disseminated at scale. These risks highlight the urgent need for reliable methods to distinguish between human- and machine-generated text.

Despite growing interest in this problem since the release of models such as GPT-2, text forensics remains less developed than its counterparts in image and video analysis. Existing approaches suffer from two key limitations. First, many methods generalize poorly across different LLMs and domains, leading to inconsistent performance. Second, most detection systems rely on binary classification, which often fails to capture the subtle generative artifacts introduced by modern LLMs.

Detection techniques can be broadly grouped into three categories. Statistical methods (e.g., GPT-2, Grover, GLTR) identify distributional irregularities in token probabilities. Watermarking approaches embed detectable signals during text generation but require control over the producing model. More recently, rewriting-based methods such as DetectGPT, RAIDAR, and SimLLM exploit differences in how LLMs paraphrase human versus synthetic text, showing strong robustness across models and domains.

However, little attention has been paid to low-resourced languages, leaving a critical gap in the global applicability of detection research. In particular, Bulgarian–a morphologically rich language with growing exposure to LLM applications–lacks dedicated detection methodologies.

In this paper, we address this gap by introducing BuST (Bulgarian Siamese Transformer), a novel rewriting-based framework for detecting machine-generated Bulgarian text. Inspired by RAIDAR's paraphrase-based detection strategy, BuST leverages a Siamese Transformer architecture to measure similarity between original and rewritten sentences, capturing subtle differences in how LLMs and humans produce text.

Our main contributions are threefold:

1. We present the first dedicated study of machine-generated text detection for Bulgarian, an underexplored low-resource language.
2. We introduce BuST, a Siamese Transformer

Proceedings of the First Interdisciplinary Workshop on OMMM associated with RANLP 2025,
pages 45–52, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/978-954-452-101-1-005

approach tailored for rewriting-based detection.

3. We evaluate BuST on a newly curated Bulgarian dataset, demonstrating its effectiveness compared to existing baselines.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 describes the dataset used. Section 4 outlines the methodology. Section 5 reports experimental results. Section 6 concludes with key findings and directions for future research.

2 Related work

This study represents a continuation of ongoing efforts within the research framework to combat the proliferation of AI-generated disinformation and synthetic media in Bulgarian. The BuSTv2 dataset, partially introduced in earlier publications, initially served as the basis for experiments centered on BERT-based fine-tuning for binary classification of human- versus machine-written texts. Earlier iterations of the dataset included a more limited number of samples and were primarily focused on traditional supervised learning approaches. In contrast, the present work introduces an expanded version of the dataset and explores fundamentally different methodological directions.

This study is inspired by the approach proposed in (Mao et al., 2024), which detects machine-generated content through paraphrastic rewriting, the current methodology shifts emphasis from direct classification toward the measurement of invariance under transformation. While this approach does not rely on model-specific watermarking or statistical fingerprinting, it proves highly effective in distinguishing AI-generated content by exploiting latent patterns of linguistic preference inherent to large language models.

In recent years, researchers have made significant progress in detecting machine-generated text, with a particularly promising direction emerging around methods that involve rewriting or perturbing the input. Unlike earlier approaches that focused on statistical measures like entropy and perplexity (Gehrmann et al., 2019; Chakraborty et al., 2023; Ghosal et al., 2023), or those that relied on syntactic and stylistic features for classification (Fröhling and Zubiaga, 2021; Nitu and Dascălu, 2024), this new wave of techniques looks at how text behaves when modified. Some systems also build on fine-tuned models like BERT or RoBERTa for simple binary classification (Maloyan et al., 2022; Bahad et al., 2024).

Rewriting-based approaches, however, take a different and increasingly effective path. They are grounded in the observation that large language models (LLMs) treat human-written and AI-generated texts differently when asked to rewrite them. For example, DetectGPT identifies machine-generated text by introducing small changes to the original and measuring how the model's confidence, or log-probability, drops. These drops tend to be more pronounced when the original was macine-generated (Mitchell et al., 2023; Xiong et al., 2024).

Another method, DetectLLM-NPR, builds on a similar idea by applying subtle perturbations and tracking how the rank of the text shifts. AI-generated content tends to react more strongly, making it easier to flag (Su et al., 2023). RAIDAR (Mao et al., 2024) takes things further by comparing how much an LLM rewrites a piece of text. It turns out that LLMs are more likely to make significant edits to human-written content—perhaps because they "perceive" it as needing more improvement—while leaving AI-generated text mostly unchanged. This difference can be captured using simple edit-distance calculations (Kavathekar et al., 2024; Zou et al., 2025), and the method has shown strong performance across various types of content.

Other systems build on the same logic. Sim-LLM, for instance, uses LLMs to generate several rewritten versions of the same text and then checks how close these are to the original to infer its origin (Nguyen-Son et al., 2024; Zou et al., 2025). Similarly, Zhu et al. (2023) show that ChatGPT tends to revise machine-generated text less than it does human-authored material.

Complementing these methodological advances, several datasets have been introduced to support detection research, particularly in Bulgarian. The M4 benchmark dataset (Wang et al., 2024) provides both scale and parallelism, with 94,000 non-parallel human-authored news articles and 9,000 parallel texts (3,000 human-written and 6,000 machine-generated) created using `davinci-003` and ChatGPT. As a multilingual dataset, it supports evaluation on both monolingual and cross-lingual detection tasks. Similarly, the MultiSocial dataset (Macko et al., 2025) collects 20,378 short texts from Telegram (9,889), Twitter (10,297),

and Gab (192), enabling the study of detection methods across social media platforms and multiple languages. In contrast, the Deepfake-BG2 dataset (Temnikova et al., 2023) is monolingual, comprising 9,824 posts from Telegram and Facebook groups evenly split between human-written and machine-generated content, with the latter produced using GPT-WEB-BG (a GPT-2 variant fine-tuned for Bulgarian) and ChatGPT, focusing on COVID-19 discourse. Collectively, these datasets expand the empirical foundation for rewriting-based detection methods and underscore the importance of cross-lingual, domain-specific, and platform-aware evaluation in Bulgarian NLP.

3 Data

To support our experiments, we compiled a dataset consisting of both formal and informal text sources: news articles and social media posts. The news article data were drawn from a publicly available Bulgarian news dataset `clickbait_news_bg`[1], while the social media texts were sampled from the dataset proposed by Temnikova et al. (2023).

We first sampled 1,623 human-written news articles, ensuring a balanced selection by drawing from different time periods and news sources. This approach aimed to capture a diverse range of topics, writing styles, and publication contexts. To obtain corresponding machine-generated samples, we used the GPT-4o-mini model to generate one synthetic version for each article, resulting in 1,623 generated texts. The final news dataset thus consists of 3,246 articles — equally divided between human-written and machine-generated content.

For the social media dataset, we randomly selected 1,000 texts — 500 written by humans and 500 generated by ChatGPT. These texts reflect more informal language and structure, offering a useful contrast to the news article domain.

We then combined the news and social media data to form a unified dataset comprising 4,246 samples in total.

3.1 Text Paraphrasing Procedure

For the purposes of our experiments, each text (regardless of its origin) was paraphrased using GPT-4o-mini. This resulted in a pair of texts for every original entry: the input text x and its paraphrased version x'. These paraphrased pairs were essential for computing text similarity, which forms the basis of our classification approach.

Different paraphrasing prompts were used depending on the source domain. For news articles, we employed the following prompt:

***Role**: 'You are a Bulgarian reporter or journalist'*
rewrite or paraphrase the text
Use Bulgarian language
Fallback message: Return "-" in case if you can not write
text

For social media texts, we used a prompt tailored to informal language:

***Role**: 'You are a Bulgarian user of social app like twitter or telegram'*
rewrite or paraphrase the text
Use Bulgarian language
Fallback message: Return "-" in case if you can not write
text

These prompts were designed to preserve the original meaning while allowing for natural variation in lexical and syntactic structure.

These (text, paraphrase) pairs were then encoded and passed into the Siamese network, which learns to detect fine-grained differences in linguistic behavior via similarity-based learning.

4 Methodology

4.1 Problem Formulation

We frame machine-generated text detection as a **binary classification** problem, where the goal is to predict a label $y \in \{0, 1\}$ for an input text x. Instead of relying solely on the raw text, we incorporate an additional predictive signal derived from *paraphrasing*. An external black-box LLM is prompted to generate a paraphrase $x' = F(p, x)$.

The hypothesis is that AI-generated text exhibits *greater semantic self-similarity* under paraphrasing than human-authored text, which typically rewrites more divergently. Thus, classification is based on both x and the semantic relationship between x and x'.

Formally, each datapoint is represented as a triple

$$(x, x', y),$$

where $y = 1$ if x is AI-generated and $y = 0$ otherwise.

[1] https://huggingface.co/datasets/community-datasets/clickbait_news_bg

4.2 Input Data

All texts are lowercased and tokenized using a WordPiece tokenizer, truncated or padded to a maximum length of 512 tokens. The dataset is split into training (60%), validation (20%), and test (20%) sets, with balanced class distributions.

4.3 Model Architecture

The detection pipeline consists of three components: an **encoder**, a **Siamese similarity mechanism**, and a **classifier**.

Encoder We experiment with two encoder families:

- **Transformer encoder:** a 6-layer stack with 8 self-attention heads per layer and hidden size 768.
- **RNN encoder:** a two-layer bidirectional LSTM ($d_{\text{emb}} = 300$, $d_{\text{hid}} = 256$, dropout = 0.3) with an additive attention mechanism, producing a 512-dimensional representation.

The RNN achieves slightly higher performance on small datasets due to fewer trainable parameters and reduced risk of overfitting. However, Transformer encoders are expected to generalize better as training data increases, benefiting from efficient parallelization and faster convergence.

Each encoder maps a sentence x to a mean-pooled embedding:

$$h = f_\theta(x), \quad h' = f_\theta(x'),$$

where f_θ denotes the encoder with parameters θ.

Paraphrastic Perplexity Test (PPT) Our method builds on the intuition of RAIDAR (Mao et al., 2024), which detects AI-generated text by comparing an input with its LLM-generated rewrite using edit distance:

$$\mathcal{L}_{\text{inv}}^{\text{RAIDAR}}(x) = D_{\text{edit}}\big(F(p,x),\, x\big),$$

where D_{edit} is the Levenshtein distance (Levenshtein, 1966).

Instead of surface-level similarity, we measure proximity in embedding space. Given a shared encoder f_θ, the *Paraphrastic Perplexity Test (PPT)* is defined as:

$$\text{PPT}(x) = \Delta\big(f_\theta(x),\, f_\theta(F(p,x))\big),$$

where Δ is a Siamese distance function.

Intuitively:

- AI-generated texts are paraphrased with higher structural invariance, yielding embeddings that remain close.
- Human-authored texts are paraphrased less predictably, producing greater divergence.

Thus, PPT replaces RAIDAR's string-level edit distance with a neural similarity metric, enabling end-to-end training while retaining robustness.

Classifier To capture the relationship between h and h', we construct a combined feature vector:

$$z = [h;\, h';\, |h - h'|;\, h \odot h'],$$

where $[\,\cdot;\cdot\,]$ denotes concatenation, $|\cdot|$ the element-wise absolute difference, and $\odot$ the element-wise product.

This vector z is passed through a multi-layer perceptron (MLP) with ReLU activations and dropout $p = 0.1$, followed by a sigmoid projection to produce a probability that x is AI-generated.

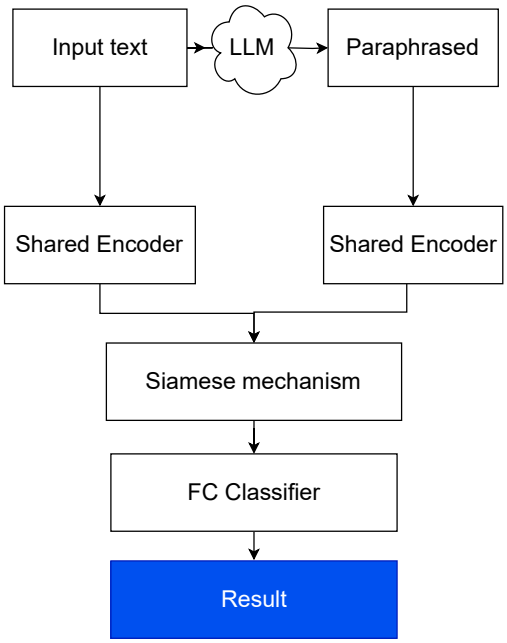

Figure 1: The BuST detection pipeline.

In summary, our pipeline first paraphrases the input text using a LLM, then encodes both the original and paraphrased versions with a shared encoder. The resulting embeddings are compared through a Siamese architecture, and the derived similarity features are passed to a classifier that predicts whether the text is human- or AI-generated (see Figure 1).

4.4 Training Configuration

Models are trained with binary cross-entropy loss, optimized using AdamW.

- BERT: learning rate 2×10^{-5}, batch size 8, weight decay 0.01, trained for 4 epochs.
- BuST: learning rate 5×10^{-5}, batch size 16, weight decay 0.01, trained for 10 epochs.
- RNN: same setup with expanded number of epochs to 25.

Both encoder families are initialized from general-domain pretraining and fine-tuned for the detection task.

4.5 Baselines

We compare our primary **Transformer-based Siamese detector with PPT** against:

1. **BERT fine-tuning:** a standard single-classifier baseline.
2. **Siamese RNNs:** with and without attention.

This tri-partite setup isolates the impact of architectural capacity and inductive bias on detection quality.

4.6 Evaluation

Models are evaluated on the held-out test set using standard classification metrics: accuracy, precision, recall, F1-score, and AUC.

To quantify the contribution of paraphrasing, we report results against two ablations:

1. A model using only the original input x.
2. A model using frozen pretrained embeddings (e.g., cosine similarity from Sentence-BERT) without fine-tuning.

5 Results

The results from our pilot experiments reveal several key insights into the performance of paraphrase-based detection models.

The Siamese models—one using a recurrent (RNN) encoder and the other based on a transformer—achieved strong performance, with accuracies above 80%. These models effectively leverage the structural and semantic similarity between original and paraphrased texts, which is central to our classification strategy. The attention-enhanced RNN performed particularly well despite the limited size of the dataset, making it a promising option for low-resource language settings like Bulgarian.

The Transformer-based Siamese model achieved the highest accuracy among the custom architectures but required more data to fine-tune effectively and exhibited greater variability across training runs.

The pretrained BERT model outperformed all other models in terms of both accuracy (93.7%) and F1 score (93.9%). Although it was not specifically optimized for paraphrase comparison, BERT proved to be a robust baseline due to its scalability and ability to generalize across diverse text types.

In terms of dataset domains, news articles were easier for models to classify correctly. This is likely due to their more formal and consistent structure, which provides clearer patterns for distinguishing human- and machine-generated text. By contrast, shorter and less structured texts—such as those from social media—led to more frequent classification errors due to fewer linguistic cues.

These findings suggest that paraphrase-based similarity learning is a viable and effective strategy for detecting machine-generated text, even in under-resourced languages. They also highlight the importance of model selection and input domain in determining detection performance.

6 Conclusion and Future Work

In this study, we introduced BuST, a novel approach for detecting machine-generated Bulgarian text by leveraging a paraphrase-based semantic similarity framework implemented via a Siamese Transformer architecture. Our method builds on recent rewriting-based detection insights, hypothesizing that the semantic (mis)alignment between an original text and its LLM-generated paraphrase contains informative signals indicative of its origin. Through experiments on a combined dataset of Bulgarian news articles and social media posts, we demonstrated that the proposed paraphrastic similarity mechanism effectively distinguishes human-written from machine-generated texts.

Our pilot results reveal that the Siamese models, particularly those using Transformer encoders, achieve strong classification performance, although pretrained BERT remains a competitive baseline. The use of paraphrased input pairs and similarity-based embeddings provides an interpretable and

Name	Mixed (Acc)	Media (Acc)	News (Acc)	F1
RNN	76.80%	-	-	79.40%
RNN + Attention	87.90%	67.8%	93.1%	88.00%
BuST	88.79% *	67.5%	93.9%	88.00%
BERT	93.70%	87.9%	93.0%	93.90%

Table 1: Results from pilot experiments using different architectures. *Transformer model showed high variance across runs.

flexible alternative to single-text classifiers, capable of capturing subtle stylistic and semantic differences. Additionally, our experiments highlight domain-specific challenges, with formal news articles being easier to classify than informal social media texts.

Dataset observations. The main part of the dataset consists of large texts such as news articles. The tests showed a significant decrease after adding social media posts, which are noticeably smaller: around 200 characters on average. The decrease was around 8%. The transformer-based model showed decrease from 96.0% to 88.8%.

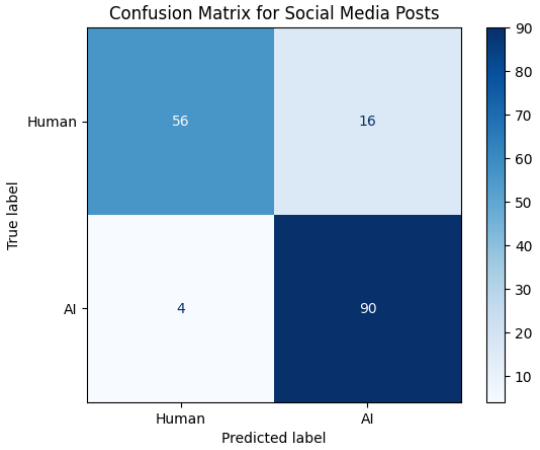

Figure 2: Confusion Matrix for Social Media Posts for BERT model

Overall, the confusion matrices (Figs. 3 and 2) provide further insight into the BERT model's performance across different text types. For news articles, the classifier correctly identified 286 out of 328 human-written texts (87.2%) and misclassified 42 (12.8%) as AI-generated, while for AI-generated articles it achieved an almost perfect accuracy, correctly classifying 288 out of 289 (99.7%) with only a single error. In contrast, the results on social media posts demonstrate reduced robustness. For human-written posts, the accuracy dropped to 56 out of 72 (77.8%), with 16 (22.2%) misclassified as AI-generated. For AI-generated posts, the classifier correctly recognized 90 out of 94 (95.7%), but still misclassified 4 (4.3%) as human-written. These results proves that while the model maintains high performance on longer, more structured texts such as news articles, it struggles with shorter and less formal social media texts, where the misclassification rate for human-written content increases significantly. This highlights the influence of text length and style on the classification accuracy of transformer-based models.

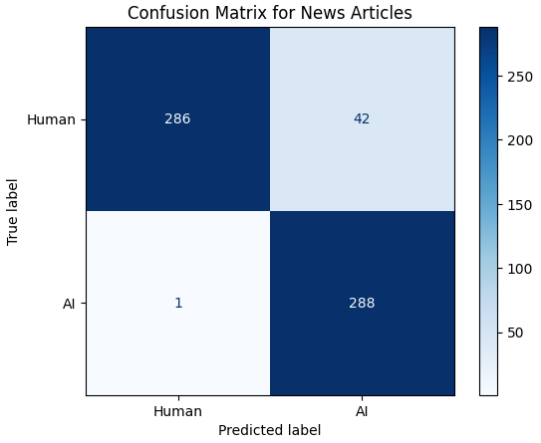

Figure 3: Confusion Matrix for News Articles for BERT model

For the BuST model, the confusion matrices (Figs. 4 and 5) reveal a less balanced performance across both text types, though with lower overall accuracy compared to BERT. On news articles, the classifier correctly identified 306 out of 318 human-written texts (96.2%), misclassifying 12 (3.8%) as AI-generated, while for AI-generated articles it achieved 268 out of 293 (91.5%) with 25 (8.5%) misclassified. In the case of social media posts, performance declined more noticeably: only 46 out of 79 human-written texts (58.2%) were correctly classified, while 33 (41.8%) were misclassified as AI-generated. For AI-generated posts, the accuracy reached 76 out of 92 (82.6%), with 16 (17.4%) incorrectly labeled as human. These findings suggest that although BuST handles longer news articles

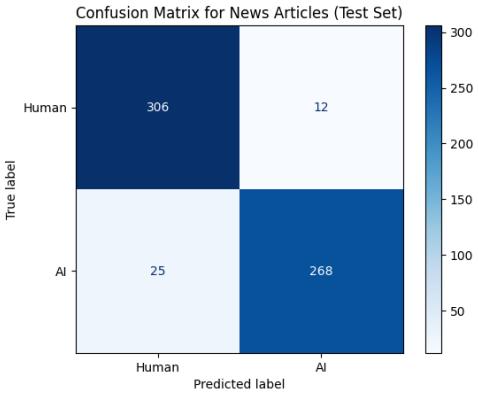

Figure 4: Confusion Matrix for News Articles for BuST model

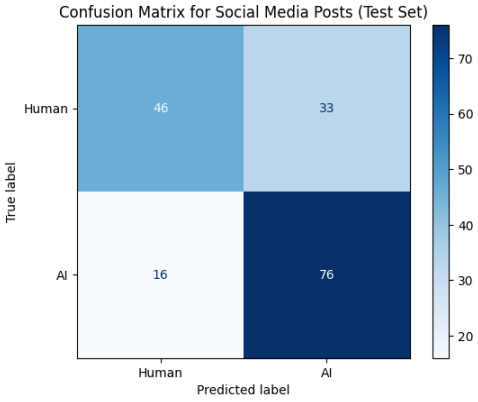

Figure 5: Confusion Matrix for Social Media Posts for BuST model

relatively well, its ability to distinguish between human and AI texts deteriorates significantly for shorter, informal content.

In comparison, while both models perform strongly on longer news articles, BuST exhibits a sharp decline in accuracy on shorter social media texts. This indicates that BuST is considerably less robust to variations in text length and style, whereas BERT maintains more stable performance across different domains.

Looking forward, several promising directions for future research emerge. First, scaling up the dataset and incorporating additional text genres and sources will be crucial to improve model robustness and generalization. Second, exploring alternative paraphrasing strategies, including diverse prompting techniques or different LLMs for generating paraphrases, may enhance the quality and informativeness of the semantic similarity signal. Third, integrating other modalities of analysis—such as stylometric features or token-level likelihoods—could complement the paraphrase similarity approach and further boost detection accuracy.

Finally, investigating model interpretability to better understand which linguistic or semantic features drive classification decisions would be valuable for practical deployment. We also envision adapting our framework to multilingual or cross-lingual settings, given the global importance of detecting synthetic text across languages.

Overall, our work contributes to the growing body of research leveraging rewriting-based signals for AI text detection and provides a foundation for developing robust detection tools tailored to under-resourced languages like Bulgarian.

Acknowledgments

This work is supported by GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no. 857155, the programme “Research, Innovation and Digitalization for Smart Transformation” 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01, and the BROD project, funded by the Digital Europe programme of the European Union under grant agreement no. 101083730.

References

Sankalp Bahad, Yash Bhaskar, and Parameswari Krishnamurthy. 2024. Fine-tuning language models for ai vs human generated text detection. *International Workshop on Semantic Evaluation*.

Souradip Chakraborty, A. S. Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *arXiv.org*.

Leon Fröhling and A. Zubiaga. 2021. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. Gltr: Statistical detection and visualization of generated text. *Annual Meeting of the Association for Computational Linguistics*.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and A. S. Bedi. 2023. Towards possibilities & impossibilities of ai-generated text detection: A survey. *arXiv.org*.

Ishan Kavathekar, Anku Rani, Ashmit Chamoli, P. Kumaraguru, Amit P. Sheth, and Amitava Das. 2024. Counter turing test (ct2): Investigating ai-generated text detection for hindi - ranking llms based on hindi ai detectability index (adihi). *Conference on Empirical Methods in Natural Language Processing*.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025. 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.

Narek Maloyan, Bulat Nutfullin, and Eugene Ilyushin. 2022. Dialog-22 ruatd generated text detection. *Computational Linguistics and Intellectual Technologies*.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. Raidar: generative ai detection via rewriting.

E. Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *International Conference on Machine Learning*.

Hoang-Quoc Nguyen-Son, Minh-Son Dao, and Koji Zettsu. 2024. Simllm: Detecting sentences generated by large language models using similarity between the generation and its re-generation. *Conference on Empirical Methods in Natural Language Processing*.

Melania Nitu and Mihai Dascălu. 2024. Beyond lexical boundaries: Llm-generated text detection for romanian digital libraries. *Future Internet*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *Conference on Empirical Methods in Natural Language Processing*.

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.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Feng Xiong, Thanet Markchom, Ziwei Zheng, Subin Jung, Varun Ojha, and Huizhi Liang. 2024. Fine-tuning large language models for multigenerator, multidomain, and multilingual machine-generated text detection. *arXiv.org*.

Biru Zhu, Lifan Yuan, Ganqu Cui, Yangyi Chen, Chong Fu, Bingxiang He, Yangdong Deng, Zhiyuan Liu, Maosong Sun, and Ming Gu. 2023. Beat llms at their own game: Zero-shot llm-generated text detection via querying chatgpt. *Conference on Empirical Methods in Natural Language Processing*.

Yueying Zou, Peipei Li, Zekun Li, Huaibo Huang, Xing Cui, Xuannan Liu, Chenghanyu Zhang, and Ran He. 2025. Survey on ai-generated media detection: From non-mllm to mllm. *arXiv.org*.