Zero-Shot Detection of AI-Generated Bulgarian Text via Binoculars Framework

Yolina Petrova
New Bulgarian University, Identrics
yolina.petrovaa@gmail.com

Nikola Blajev
Identrics, Experian
nikola.blajev@msn.com

Abstract

We present a zero-shot method for detecting AI-generated text in Bulgarian using a binoculars framework. The approach leverages two closely related models to compute a score based on the ratio of perplexity to cross-perplexity, requiring no task-specific training. Evaluated on a small curated dataset (367 calibration and 50 evaluation samples), the method achieves F1=0.91 (AUC=0.97) on calibration data and F1=0.99 on evaluation data, indicating strong separability between human-written and AI-generated content. We discuss limitations related to dataset size and generalization. Our findings suggest binocular LLM scoring is a promising lightweight approach for low-resource languages.

Keywords: large language models, synthetic content detection, perplexity-based metric.

1 Introduction

The emergence of Large Language Models (LLMs) has fundamentally transformed how textual content is produced and consumed, enabling the large-scale generation of coherent and contextually relevant text. However, this capability also amplifies the risk of misinformation, as such models can produce persuasive yet factually incorrect or misleading content that is difficult to distinguish from human-written text. One of the primary challenges in this domain is the detection of AI-generated (synthetic) content. Identifying whether a piece of text is synthetic is a crucial first step in broader analytical pipelines, including propaganda detection, stance classification, and influence analysis.

Existing approaches to AI-generated text detection fall broadly into three categories: watermarking-based methods, supervised classifiers, and statistical detection techniques. Watermarking approaches embed detectable signals directly into generated text (Kirchenbauer et al., 2023); however, they require control over the generation process and are not applicable to already generated or third-party content. Supervised approaches train classifiers to distinguish between human-written and AI-generated text (Wang et al., 2023). While effective, these methods depend on large labeled datasets and may suffer from domain shift.

Statistical detection methods, including curvature-based approaches such as DetectGPT (Mitchell et al., 2023), aim to identify intrinsic differences between human and AI-generated text without requiring explicit supervision. More recently, zero-shot methods such as Binoculars (Hans et al., 2024) have demonstrated strong performance by leveraging differences between language models, without task-specific training.

However, most existing work focuses on high-resource languages, and the applicability of these methods to low-resource settings remains underexplored. In particular, languages such as Bulgarian lack large-scale annotated datasets and systematically evaluated detection pipelines, making supervised approaches difficult to deploy and motivating the need for lightweight, training-free alternatives.

Contributions

- We propose the first adaptation of the Binoculars framework to Bulgarian, a low-resource language, and analyze its effectiveness on synthetic content detection without task-specific training.
- We provide an empirical analysis of binocular score distributions in a low-resource language setting.
- We introduce a calibrated thresholding strategy tailored to small datasets.

2 Method

2.1 Dataset

To evaluate the proposed approach in a low-resource setting, we use a subset of the publicly available WASPer dataset (Petrova et al., 2025). While the dataset was originally developed for multilingual social media propaganda detection in Bulgarian and English, it provides a suitable basis for studying AI-generated content detection in Bulgarian, where annotated resources remain limited.

2.1.1 Calibration Dataset

The calibration dataset consists of a balanced set of examples including both **organic (human-written)** and **synthetic (AI-generated)** content.

- **Human-written examples** were manually selected from comment sections under variety online news sources to ensure diversity in topics, style, and language use. The samples were reviewed by two independent annotators to ensure high quality and representativeness of real-world Bulgarian text.
- **AI-generated examples** were produced using large language models under controlled prompting conditions. For Bulgarian, BgGPT models (Alexandrov et al., 2024) were employed. The synthetic samples were generated using **few-shot prompting** where the models were given representative human-written examples as input, allowing the generation of realistic and varied text. To ensure diversity, the prompts were designed to vary across: topics (e.g., politics, social issues), style (formal/informal) and sentiment (positive/negative)[1]. All generated content was subsequently manually reviewed for quality and consistency.

The final calibration dataset for Bulgarian that we used for the current study consists of 367 examples, including human-written and AI-generated samples. While the dataset size is relatively small, it is sufficient for evaluating threshold-based zero-shot methods, as no model parameters are learned and the calibration set is used solely for estimating score distributions.

2.1.2 Evaluation Dataset

A separate evaluation set of 50 examples (26 human-written (organic) texts and 24 AI-generated (synthetic) texts) was curated to assess the performance of the AI-generated content detection pipeline. This evaluation dataset was sourced similarly to the calibration set, ensuring a representative mix of human-written and AI-generated texts across topics.

The texts in both the calibration and evaluation datasets are relatively short, with an average length of approximately 1.5 sentences. The use of short, user-generated texts reflects real-world scenarios such as social media comments, where detecting AI-generated content is particularly challenging due to limited context. Importantly, no task-specific fine-tuning over this dataset was performed. The distribution of human-written versus AI-generated samples in calibration and evaluation datasets is summarized in Table 1.

Dataset Split	Human	AI-generated
Calibration	206	161
Evaluation	26	24

Table 1: Distribution of human-written (organic) and AI-generated (synthetic) texts in the Bulgarian calibration and evaluation dataset.

2.2 Binoculars Scoring Framework

We developed a zero-shot binary classifier using the Binoculars framework as a backbone. The classifier relies on *perplexity*, a standard measure of how well a language model predicts a given text. The method analyzes each input using two LLMs, referred to as the *performer* and the *observer*, sharing the same tokenizer. AI-generated text tends to occupy regions of the probability space that are consistent across similar models, while human-written text shows larger discrepancies. However, perplexity alone is not a reliable indicator of AI-generated content, as it is also sensitive to linguistic complexity and unusual inputs. Therefore, we use two LLMs, referred to as the "performer" and the "observer," and we introduce the binocular score: ratio of perplexity (calculated from the “performer” model) to cross-perplexity, a notion of how surprising the next token prediction of one model ("predictor") is to another model (“observer”). The cross-perplexity term effectively “calibrates” the score for high-perplexity prompts.

[1]To ensure that the dataset covers a wide range of topics without bias **topic modeling** using BERTopic (Grootendorst, 2022) was employed, a state-of-the-art library that leverages language model embeddings to identify coherent topics in text data. This process groups texts into distinct topic clusters, enabling balanced sampling of both human and AI-generated examples across topics.

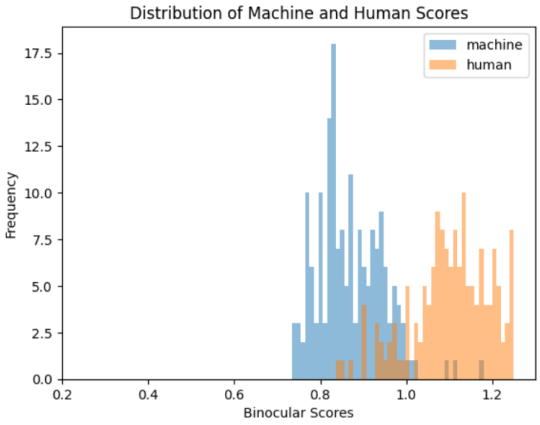

Figure 1: Distribution of Binocular scores for Human-written and AI-generated Texts

Figure 1 illustrates the distribution of binocular scores. AI-generated text exhibits low relative divergence (similar perplexity (PPL) and cross-perplexity (XPPL)), whereas human-written text exhibits higher relative divergence.

As a result:

- AI-generated text → similar PPL and XPPL → low relative divergence
- Human-written text → larger discrepancy between PPL and XPPL → high relative divergence

This makes the binocular score a potential robust discriminator, particularly in zero-shot settings.

2.2.1 Mathematical Formulation

Formally, the method computes:

- **Perplexity**: measures how well the performer model predicts the input text.

$$\log \mathrm{PPL}_M(s) = -\frac{1}{L}\sum_{i=1}^{L} \log\left(Y_i(x_i)\right) \quad (1)$$

  where:
  - M is the performer language model;
  - s is the input sentence;
  - L is the number of tokens in s;
  - $Y_i = M(s)$ represents the output probability distribution of the model at position i, i.e., the probability that word j of the vocabulary will follow the string up to word i;
  - x_i is the i-th word in s.

- **Cross-Perplexity**: measures how well the observer model predicts tokens generated under the performer model's distribution.

$$\log \text{X-PPL}_{M_1,M_2}(s) = -\frac{1}{L}\sum_{i=1}^{L}\left(M_1(s)_i \cdot \log(M_2(s)_i)\right) \quad (2)$$

  where:
  - M_1 and M_2 are the performer and observer models, respectively;
  - s is the input string of characters;
  - L is the number of tokens in s.

The actual **binocular score** is defined as the ratio between the two:

$$B_{M_1,M_2}(s) = \frac{\log \mathrm{PPL}_M(s)}{\log \text{X-PPL}_{M_1,M_2}(s)} \quad (3)$$

This normalized ratio captures how likely a text is AI-generated while accounting for linguistic complexity and model-specific variability.

2.3 Model Selection and Algorithm Configuration

We evaluated different permutations of Mistral and BgGPT as candidates for the performer and observer models. We selected those particular model families due to their lightweight architecture, strong performance and open-source availability. After thorough testing, we chose the following configuration, as it yields the best performance when measured with the area under the ROC curve (AUC) metric:

- Performer: `BgGPT-7B-Instruct-v0.1`
- Observer: `BgGPT-7B-Instruct-v0.2`

The performer-observer configuration was selected to maximize agreement on AI-generated texts while differentiating human-written texts, measured via AUC.

3 Results

3.1 Threshold Selection

To classify a text as human-written or AI-generated, we compute the binocular score $B(s)$ for each input sample s. A threshold τ is then applied such that:

$$\hat{y}(s) = \begin{cases} \text{AI-generated}, & \text{if } B(s) \leq \tau \\ \text{Human-written}, & \text{otherwise} \end{cases} \quad (4)$$

We selected τ based on the calibration set to achieve a target **true positive rate (TPR)** of 90%, which balances the detection of AI-generated content with the risk of false positives. The choice of 90% TPR reflects a practical trade-off seen on Figure 2: we prioritize detecting most AI-generated texts while keeping the false positive rate (FPR) at an acceptable level.

Once the threshold is fixed, we compute the standard precision, recall, and F1-score as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{Recall (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

where:

- TP = number of AI-generated texts correctly classified
- FP = number of human-written texts incorrectly classified as AI-generated
- FN = number of AI-generated texts incorrectly classified as human-written

By tuning τ on the calibration set, we ensure that the F1-score computed on the evaluation set reflects a balanced performance between precision and recall. In this case for Bulgarian, a threshold of $\tau = 0.968$ resulted in an F1-score of 0.905.

3.2 Evaluation on Unseen Data

The F1 score of the classifier was further tested on the evaluation dataset, consisting of 26 organic and 24 synthetic comments. On the evaluation dataset, the F1-score achieved 0.99. These results indicate strong generalization, particularly for Bulgarian. However, given the small size of the evaluation dataset, these results should be interpreted with caution. Overall, the binocular score distributions show clear separation between human and AI-generated samples, with minimal overlap. This demonstrates the effectiveness of binocular scoring in low-resource settings, even with small calibration sets.

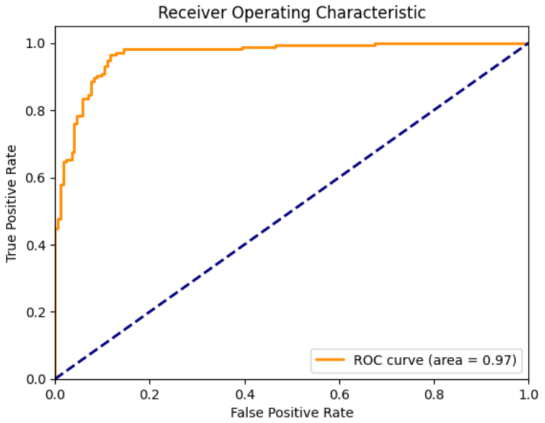

Figure 2: ROC curve for Bulgarian dataset

3.3 Error Analysis

We observed that misclassifications primarily occur in: (i) short texts with limited contextual information, i.e., ”това е вярно”; (ii) AI-generated text that has been manually edited, and (iii) linguistically ambiguous or stylistically neutral content.

4 Discussion

We present a training-free approach for detecting AI-generated text using a binocular scoring mechanism. The method achieved strong performance on Bulgarian datasets while remaining lightweight and easy to deploy. Our findings suggest that statistical detection methods based on model agreement provide a scalable solution for identifying synthetic content, particularly in low-resource languages.

However, several limitations remain: (i) performance may degrade for highly edited or paraphrased AI-generated text, (ii) the method depends on the availability of suitable model pairs, and (iii) cross-lingual generalization requires further investigation. Additionally, the synthetic data generation process may introduce biases related to prompting strategies and model-specific characteristics. Future work will focus on addressing these limitations in several directions. First, we plan to evaluate the method on substantially larger and more diverse Bulgarian corpora, including texts from different domains, writing styles, and LLM families, to improve the robustness of threshold estimation and assess generalization. Second, the detector will be evaluated on paraphrased and post-edited AI-generated texts, potentially by combining binocular scoring with complementary linguistic features to improve robustness against human editing. Third, we intend to investigate alternative performer–observer

model pairs, including multilingual and more recent Bulgarian language models, to better understand the impact of model selection on detection performance. Finally, extending the approach to additional low-resource languages and studying cross-lingual calibration strategies will help determine whether binocular scoring can serve as a language-independent framework for synthetic text detection.

Despite these limitations, the approach provides a practical solution for AI-generated content detection, which could act as a filtering step to pipelines for misinformation and propaganda detection systems. The proposed method can be integrated as a lightweight filtering component in real-time moderation pipelines.

References

Anton Alexandrov, Veselin Raychev, Dimitar I. Dimitrov, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. BgGPT 1.0: Extending English-centric LLMs to other languages.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

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. 235:17519–17537.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Yolina Petrova, Todor Kiryakov, Devora Kotseva, and Boryana Kostadinova. 2025. WASPER: A Bulgarian-language model for detecting propaganda in social media content. In *Papers from the International Scientific Conference of the European Studies Department, Jean Monnet Centre of Excellence, Faculty of Philosophy at Sofia University "St. Kliment Ohridski"*, volume 12, pages 223–230.

Yuxia Wang et al. 2023. ChatGPT is on the horizon: Could AI-generated text replace human-written text? *arXiv preprint arXiv:2303.11130*.