Detecting AI-Generated Bulgarian Text: A Two-Step Multi-Class Classification Approach

Boyan Bogdanov
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
boyan.bogdan@gmail.com

Daniel Georgiev
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
georgiev.daniel00@gmail.com

Dimitar Dimitrov
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
ilijanovd@fmi.uni-sofia.bg

Ivan Koychev
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
koychev@fmi.uni-sofia.bg

Preslav Nakov
Mohamed bin Zayed University of
Artificial Intelligence
Abu Dhabi, UAE
preslav.nakov@gmail.com

This paper introduces a novel two-step multi-class classification system to identify varying degrees of machine involvement in Bulgarian text. As Large Language Models (LLMs) proliferate, distinguishing original human writing from machine-assisted or machine-generated content is crucial to prevent misinformation and preserve educational integrity. We developed a comprehensive dataset in Bulgarian encompassing purely human-written texts and four distinct mixed-content categories, such as machine-continued and machine-polished text. Our proposed pipeline utilises a binary ensemble classifier combining stylometric features, XLM-RoBERTa, and an adapted Binoculars model to first filter out human-written text. Subsequently, a specialised multi-class Support Vector Machine categorises the remaining machine-involved texts. Our findings indicate that this hierarchical approach improves the reliable identification of human authorship and enhances the overall accuracy of multi-class text detection compared to single-step methods.

Keywords: AI-generated text detection, large language models, benchmark evaluation, Bulgarian language

1. Introduction

The rapid spread of sophisticated Large Language Models (LLMs) has made it increasingly difficult to distinguish between text written by humans and text generated by machines, posing substantial problems in many areas. When LLM-generated content is passed off as

https://doi.org/10.47810/JCLIB.2.2026.01

original human work, it can fuel the spread of false information, violate intellectual property rights, and lower educational quality. Furthermore, the inability to reliably detect machine-generated text risks contaminating public datasets, which in turn degrades the quality of training data for future language models and hinders downstream linguistic research.

To tackle this, our paper introduces a new system designed to classify Bulgarian text into multiple categories. We built an extensive collection of texts that includes purely human-written content, purely machine-generated content, and four different types of mixed content: text that started with human input and was finished by a machine, text written by humans and refined by a machine, text generated by a machine but then made to sound more human, and text produced from deeply mixed human and machine involvement.

Our study tests both established machine learning techniques and newer LLM-based classifiers using this unique, multi-category dataset. Our aim is to create a reliable standard for identifying different levels of machine involvement in Bulgarian text, moving beyond simple yes/no detection to offer a more precise and context-sensitive analysis.

In summary, our main contributions are:

- The creation of a novel, multi-category Bulgarian dataset featuring purely human, machine-generated, and deeply interwoven text classes.
- An evaluation of the limitations of single-step multi-class models in distinguishing authentic human text from machine-assisted text.
- The proposal of a high-performing, two-step hierarchical classification pipeline that significantly improves the recall of human-written Bulgarian text.

2. Related work

The detection of machine-generated text is deeply rooted in traditional authorship attribution and stylometry, where machine learning techniques established the foundational methodologies. Early research heavily relied on feature-based methods such as n-gram frequencies, TF-IDF representations, and stylometric markers, features demonstrated to effectively distinguish writing styles (Stamatatos 2009), fed into classic classifiers like Support Vector Machines (SVMs), decision trees, and random forests (Joachims 1998; Potthast et al. 2017; Kestemont et al. 2019). These conventional models demonstrated significant success in identifying authorial style and, in some cases, even outperformed early neural networks in identifying the specific Large Language Model (LLM) used for generation. As the field evolved toward binary classification (human vs. machine) of increasingly sophisticated generated text, fine-tuned transformer models such as RoBERTa have consistently shown superior performance (Uchendu et al. 2020).

A significant factor influencing detection difficulty is the scale of the generator model; text produced by larger, more complex LLMs is inherently more challenging to identify (Jawahar, Abdul-Mageed, and Lakshmanan 2020; Solaiman et al. 2019; Crothers, Japkowicz, and Viktor 2023). While detectors trained on output from smaller models show moderate success in identifying text from larger ones, the inverse is more effective. Feature-based methods relying on stylometry and linguistic properties like text perplexity remain relevant,

but typically require longer text samples and are sensitive to the generator's configuration settings (Crothers, Japkowicz, and Viktor 2023).

Modern detection efforts heavily leverage transformer-based architectures. The 2024 PAN Voight-Kampff Generative AI Authorship Verification Task (Bevendorff et al. 2024) saw a majority of successful systems using embeddings from models like BERT, DeBERTa, and RoBERTa, often feeding these features into classifiers like LSTMs or SVMs (Bevendorff et al. 2024). The top-performing solution employed an ensemble of fine-tuned LLMs and the zero-shot detector Binoculars (Hans et al. 2024), which relies on a comparison between two LLMs. This highlights a trend towards combining multiple advanced models for state-of-the-art results.

However, many of these advanced detectors exhibit significant performance degradation when tested on out-of-domain data (Wang et al. 2024; Tufts, Zhao, and Li 2025). Zero-shot models such as Grover (Zellers et al. 2019) are effective primarily within the specific domain for which they were designed, such as news articles. This domain dependency is a recurring challenge, with some research exploring adversarial training (e.g., DANN) to create more domain-agnostic models (Abassy et al. 2024).

The task of multi-class classification, which mirrors the work in this paper, was recently defined by the 2025 PAN Voight-Kampff Generative AI Detection shared task (Bevendorff et al. 2025). The results from this task, as shown in Table 3 of the cited paper, indicate that fine-tuned, large-scale models like DeBERTa-v3-Large and Qwen-4B are favored by the top-performing teams (Macko 2025; Li, Qi, and Yan 2025; Zheng et al. 2025). Many successful approaches involve sophisticated fine-tuning strategies, including data augmentation for underrepresented classes, multi-head attention mechanisms, and Mixture of Experts (MoE) architectures (Li, Qi, and Yan 2025; Teja, Yadagiri, and Pakray 2025; Voznyuk, Gritsai, and Grabovoy 2025).

While many recent top-performing systems employ single-step architectures, the concept of a two-step or hierarchical approach is also an established approach in machine generated text detection. Previous studies have utilized multi-stage pipelines to first separate human-written text from generated content before applying more granular categorization, such as identifying the specific generator model or pinpointing the exact boundary where human authorship transitions into machine continuation (Kushnareva et al. 2024; Zeng et al. 2024). We attempt to build upon this hierarchical framework, adapting it to the multi-class setup containing deeply interwoven, machine-polished, and machine-humanised texts, which remains an underexplored task for low-resourced languages such as Bulgarian.

Research on Bulgarian text detection reveals unique challenges and is an actively developing field. Early efforts in identifying machine-generated Bulgarian content on social media demonstrated the difficulty of distinguishing textual deepfakes from human discourse, highlighting the need for robust, language-specific tools (Temnikova et al. 2023). More recently, models tailored specifically for Bulgarian, such as the Siamese Transformer model BuST, have been proposed to address these nuances (Maslo and Gargova 2025). Furthermore, a study using the M4 dataset found that detectors for Bulgarian perform best when trained exclusively on Bulgarian data, even in cross-generator scenarios (Wang et al. 2024).

The multi-class detection task is complicated further by the specific characteristics of the Bulgarian language and the varying quality of generated texts. Research shows

that modern multilingual LLMs often rely on English-centric vector spaces, leading to outputs that resemble literal translations and exhibit syntactic calquing (Li et al. 2025). For Bulgarian specifically, studies have observed that models sometimes hallucinate by introducing vocabulary and grammatical structures from more widely resourced Slavic languages, such as Russian (Berbatova and Salambashev 2023). However, the contemporary models utilized in our dataset (e.g., Gemini 2.5 Flash, o4-mini) often generate fluent and grammatically correct text. Consequently, the detection task has become significantly more difficult as classifiers can no longer rely solely on superficial grammatical mistakes or signs of literal translation. Instead, they must capture subtle stylistic rigidities, repetitive phrasing, and the unnatural blending of styles that occurs when human and machine inputs are interwoven, making multi-class detection a non-trivial problem.

Moreover, the problem is uniquely challenging for Bulgarian due to significant resource constraints. Unlike English and other high-resource languages, which benefit from numerous large-scale datasets specifically for machine-generated text detection, Bulgarian lacks comparable annotated corpora for multi-class classification. The majority of existing machine-generated text detection datasets are English-centric, and adapting these resources to Bulgarian is non-trivial due to linguistic differences.

3. Dataset

In this section, we describe in details how we construct our multi-class Bulgarian dataset. We first describe the sources and extraction processes for compiling the original human baseline. Subsequently, we outline the specific prompting strategies and large language models employed to generate the four distinct classes of machine-involved content. Finally, we discuss the text preprocessing steps required to prepare this diverse corpus for the classification pipeline.

3.1 Human-written texts

To ensure a diverse collection of human-written texts, we selected several sources to guarantee sufficient variation in content properties such as domain, author, topic, and length. We then used a portion of this collected data to generate the four other classes of text, employing multiple Large Language Models for each text.

We chose the online library of New Bulgarian University (NBU),[1] as a primary source for human-written texts due to its large volume of published content and relative accessibility enabling filtering texts by various fields.

As a first step, we filtered documents by their publication year, with a cutoff date of 2020. We did this to avoid contaminating the human-written dataset with texts that could have potentially been created using LLMs. To simplify the text extraction process, we further narrowed down the selection to only PDF files, excluding other materials like audio-visual content, entire books, and presentations. The library's search engine paginates the results

1 https://eprints.nbu.bg/

but also provides a JSON file containing hyperlinks to the articles, along with their titles, unique identifiers, and other metadata.

Class	Illustrative Snippet
Human-written	Като общо правило, материята на публичното право постоянно повдига множество практически и теоретични въпроси. Причините за това могат да се търсят в няколко насоки... *As a general rule, the subject matter of public law constantly raises numerous practical and theoretical questions. The reasons for this can be sought in several directions...*
Machine-continued	Като общо правило, материята на публичното право постоянно повдига множество практически и теоретични въпроси. Причините за това могат да се търсят в няколко насоки. Първо, сложността на правната рамка често води до различни интерпретации... *As a general rule, the subject matter of public law constantly raises numerous practical and theoretical questions. The reasons for this can be sought in several directions. First, the complexity of the legal framework often leads to different interpretations...*
Machine-polished	Материята на публичното право постоянно повдига множество практически и теоретични въпроси поради няколко причини. Първо, динамиката на политическите... *The subject matter of public law constantly raises numerous practical and theoretical questions for several reasons. First, the dynamics of political...*
Machine-humanised	Правната доктрина и практика постоянно се стремят да дефинират и обвържат основните категории, които формират публичното право. В центъра на това... *Legal doctrine and practice constantly strive to define and connect the main categories that form public law. At the center of this...*
Deeply-mixed	Като общо правило, материята на публичното право постоянно повдига множество практически и теоретични въпроси. Това се дължи на многообразието на публичноправните субекти... На първо място, динамиката на... *As a general rule, the subject matter of public law constantly raises numerous practical and theoretical questions. This is due to the diversity of public law subjects... First, the dynamics of...*

Table 1
A condensed structural illustration of the five text classes based on a single source paragraph. Human text is highlighted in grey, while machine-generated text is in green.

The filtered articles, typically several pages long, were unsuitable for the detection task, as real-world scenarios rarely involve generating such long texts at once. Therefore, we segmented each article into paragraphs.

Next, we extracted the text from each PDF document. We experimented with several specialized Python libraries, including spacy-layout, pdfminer, PyMuPDF, and PyPDF. However, due to the non-standard format of the articles, these open-source libraries failed to produce satisfactory results. Common issues included the inability to correctly identify paragraph breaks, improper spacing (either missing or excessive), and scrambled lines that rendered the text meaningless. The variety and inconsistency of these errors made it difficult to develop a common data cleanup solution.

We found the most effective system for text extraction to be Azure AI Document Intelligence, a paid cloud service accessible via a Python library.[2] It uses AI to perform tasks such as document classification, table extraction, and identifying document structure, including paragraphs and titles.

In addition to the academic articles from the New Bulgarian University library, we used several other corpora of human-written texts to study the task with a more extensive dataset. Relying on texts from limited domains can lead to classifiers overfitting to these domains and performing poorly on unseen data. For this reason, we chose to include two additional domains in our dataset: news articles and Wikipedia pages, as these domains are also vulnerable to abuse of LLM involvement and offer more variety in text structure and traits. Part of the news articles are reused from the SemEval 2025 Task 10: Multilingual Characterization and Extraction of Narratives from Online News (Rosenthal et al. 2025), while the rest of the news dataset was gathered following the data collection practices described in the SemEval shared task. The Wikipedia pages subset was built using their public API which allows exporting pages into a plain text format.

We followed the same procedure for generating the other text classes from this dataset as we did for the academic articles. Due to the different properties of these new texts, being shorter and more self-contained on average, we adjusted the instructions for the LLMs to improve the text generation quality, e.g. by better reflecting the expected traits of the output.

3.2 Machine-involved texts

For the creation of the text corpus for classes requiring a Large Language Model, we used several contemporary multilingual models: Gemini 2.0 Flash, Gemini 2.5 Flash, OpenAI o4-mini, and BgGPT 2.

The **machine-generated** class is not part of the classification task's definition, but it's required as the basis for the **machine-generated, then machine-humanised** class. We generated machine-generated texts by passing only the title of human-written articles, without using any of their content. To approximate the style of human articles while maintaining diversity, we specified a target number of paragraphs or words in the prompts for some examples. We then split the resulting text into paragraphs, which served as the

2 https://azure.microsoft.com/en-us/products/ai-foundry/tools/document-intelligence

basis for some of the other classes. Examples of titles and prompts used for the generation of the texts of this class can be found in the Appendices in Table 9.

For the **human-initiated, machine-continued** class, we employed two main approaches with several prompt variations. The first involved splitting a human-written paragraph at the end of a sentence to create two parts. We included the first part in the prompt to the LLM, tasking it with completing the paragraph to a length approximately equal to the original human-written one. However, a single paragraph's beginning does not always provide sufficient context for the LLM to generate a meaningful continuation that aligns with the original. To address this, we randomly prepended between 0 and 5 preceding paragraphs to provide additional context. An visualisation of how a paragraph is split into two can be found in Table 8 in the Appendices.

The prompts we used for generating the **human-written, then machine-polished** class were the most straightforward ones, representing instructions commonly used in real-world scenarios. To generate this part of the corpus, we provided a human-written paragraph to the LLM along with a simple instruction such as "improve the readability of this paragraph", "paraphrase this paragraph", or "make this paragraph clearer".

In the **machine-generated, then machine-humanised** class, the setup is similar to the human-written, then machine-polished class, with a few key differences. Most notably, an LLM originally wrote the source text instead of a human. For this purpose, we used text from the machine-generated dataset. We then instructed the generator to revise the text by adopting a persona (e.g., an assistant, a graduate, a student), attempting to simplify it or to make the text seem less like it was written by an LLM. To simulate a real-world situation where a text might be the result of several sequential instructions, we fed the output from the previous step back into the same LLM with a modified instruction in the prompt as shown in Table 10. Each paragraph randomly underwent between 0 and 2 additional LLM humanisation iterations after the initial generation.

The **deeply-mixed** texts, which contain interwoven parts of human-written and machine-generated paragraphs, resemble the human-initiated, machine-continued class. In both cases, we deleted a portion of the human-written text for the LLM to fill in. For the mixed texts, this masked portion can appear one or more times within the paragraph, removing entire sentences. To differentiate the mixed texts from the "human-initiated, machine-continued" class, we avoided masking the very beginning or end of texts. Similarly to the "human-initiated, machine-continued" class, we sometimes appended several preceding paragraphs at the start of the current one to serve as context, enabling the LLM to fill in the missing parts with more relevant and accurate content. Again, we used several different prompts to maintain diversity in the dataset. Table 11 showcases the steps in preprocessing the text in order to generate e deeply-mixed version of it.

Table 2 specifies the number of texts from the different sources in all classes. Table 3 shows the distribution per LLM used.

	Source Domain			
Text Class	**NBU**	**News**	**Wikipedia**	**Total**
Human-written	13,919	23,779	7,667	45,365
Machine-generated	23,970	36,889	9,220	70,079
Machine-continued	28,242	601	0	28,843
Machine-polished	28,319	2,850	0	31,169
Machine-humanised	17,885	1,445	0	19,330
Deeply-mixed	15,835	0	0	15,835
Total	**128,170**	**65,564**	**16,887**	**210,621**

Table 2
Distribution of texts by class and source domain.

	Language Model				
Text Class	**BgGPT-27B**	**Gemini 2.5 Flash**	**Gemini 2.0 Flash**	**OpenAI o4-mini**	**Total**
Human-written	–	–	–	–	45,365
Machine-generated	24,504	6,480	7,642	31,453	70,079
Machine-continued	9,148	9,526	601	9,568	28,843
Machine-polished	9,260	9,498	501	11,910	31,169
Machine-humanised	5,009	6,166	274	7,881	19,330
Deeply-mixed	0	7,863	0	7,972	15,835
Total	**47,921**	**39,533**	**9,018**	**68,784**	**210,621**

Table 3
Distribution of texts by class and LLM used.

3.3 Text Preprocessing and Tokenization Analysis

The raw extracted texts required specialized natural language processing to be viable for the multi-class Support Vector Machine (SVM). We utilized the `spaCy` library, configured specifically for the Bulgarian language, to perform tokenization, lemmatization, and stop-word removal.

A critical finding during our preliminary experiments was the inverse relationship between aggressive text normalization and classification accuracy for machine-generated text. Specifically, applying lemmatization slightly degraded the model's performance. The generalization of tokens achieved through lemmatization inadvertently stripped away morphological variations and specific word forms that serve as crucial stylistic markers (artifacts of generation) left by the LLMs.

Similarly, the removal of stop words presented a trade-off. While removing the approximately 1,000 stop words from the 280,000-word corpus shifted the vocabulary focus toward content-heavy words, it led to heavy overfitting during the multi-class SVM training. The

model achieved near-perfect scores on the training set but failed to generalize, indicating that LLMs often expose their nature through the structural use of highly frequent, seemingly "meaningless" stop words. Consequently, our final Bag-of-Words representation relies on simple `spaCy` tokenization without lemmatization or stop-word removal. The natural language processing is performed by using parts of the bespoke pipeline for Bulgarian developed by (Berbatova and Ivanov 2023).

4. Two-step classification pipeline

Our preliminary studies demonstrated the limitations of a single-step approach. When we trained a multi-class SVM to directly classify texts into all five categories, it struggled to reliably distinguish purely human-written content from the various forms of machine-involved text. This approach yielded a recall of 0.729 for the human-written class, as shown in the Single-Step Recall column of Table 6, indicating that a significant portion of human-authored texts were being misclassified as machine-generated or machine-assisted.

The primary motivation for developing the two-step pipeline is to address this specific shortcoming. By introducing a specialised, high-performance binary classifier in the first stage, we aim to more accurately filter out human-written content, thereby increasing the recall for this critical class and reducing the number of false positives that are unnecessarily passed to the second-stage classifier.

To address the nuanced task of identifying different forms of machine involvement in text generation, we propose a two-step classification pipeline. This hierarchical approach is designed to first distinguish purely human-written text from any text involving machine generation and then to further categorise the specific nature of the machine's contribution. The pipeline processes a given text and assigns it to one of five classes.

4.1 Data Splitting and Leakage Prevention

A fundamental principle in training and evaluating machine learning models is preventing data leakage between the training and testing sets. Because the lengthy articles from the New Bulgarian University (NBU) library were segmented into multiple individual paragraphs to match the generation constraints of LLMs, a naive randomized split introduces a severe vulnerability.

If a simple split is employed, different paragraphs from the same source article can simultaneously appear in both the training and testing subsets. This creates an artificial overlap in highly specific vocabulary, formatting, and authorial style. In a real-world environment, the system would not have prior access to adjacent paragraphs of an unseen text. To simulate a realistic operational environment and ensure objective evaluation, we implemented a "parent-child" grouping strategy during the dataset split. We enforce a strict constraint ensuring that all paragraphs ("children") originating from the same article ("parent") are grouped exclusively in a single subset, thereby preventing any stylistic or thematic data leakage.

Applying this parent-child grouping strategy for the first stage of the classification pipeline, the dataset was partitioned into training and testing subsets. The training set is made up of 112,451 samples while the test set contains 15,586 samples, maintaining an approximate 90:10 ratio.

Following this initial parent-child grouped split, the training set underwent a secondary partitioning to facilitate ensemble training. The 112,451 training samples were divided into two subsets: one for training the base models from the stacking ensemble model (approximately 96,865 samples) and another for training the meta-model (approximately 15,586 samples). By holding out a separate evaluation set, we ensure that the ensemble performance metrics reflect genuine generalization on completely unseen data, maintaining the integrity of our results and preventing the ensemble from having an unfair advantage through exposure to its training distribution. The final ratio of the three subsets is 80:10:10.

4.2 Binary ensemble

The first stage of our pipeline employs a binary classifier to perform the initial, high-level distinction between human-written and machine-involved text. This classifier is an ensemble model that leverages three distinct detection signals: an XGBoost classifier trained on stylometric text features, a fine-tuned XLM-RoBERTa instance and a version of Binoculars adapted to Bulgarian by replacing the original Falcon observer and performer LLMs with SambaLingo-Bulgarian Base and Instruct LLMs (Csaki et al. 2024), which in turn are pretrained Llama models. We evaluated these three solutions and found they complement each other, motivating us to implement a logistic regression ensemble that classifies by scaling the output of each model by a weight learned during training on a held-out subset.

Class	Precision	Recall	F1
0 (Human)	0.954	0.838	0.892
1 (Machine)	0.943	0.985	0.963
Macro Avg	**0.948**	**0.911**	**0.928**

Table 4
Performance of the first-stage binary ensemble classifier.

The performance of this first stage is shown in Table 4. If the ensemble classifies a text as human-written, it is assigned its final label, and no further processing occurs. However, if the text is flagged as machine-involved, the system passes it to the second stage of the pipeline for a more granular analysis.

4.3 Multi-class SVM

The system forwards texts identified as machine-involved by the first step to a second-stage classifier for more detailed analysis. It is important to note that this second classifier is a separate multi-class Support Vector Machine (SVM), distinct from the one used in our initial

single-step experiments. We trained this new SVM specifically on a dataset containing only the four machine-involved classes, rather than all five. The purpose of this specialised SVM is to differentiate between the four distinct categories of machine-assisted text generation.

For feature representation, the SVM utilizes a Bag-of-Words model. Each text is converted into a numerical vector that represents the frequency of words from a vocabulary built upon the training corpus. This traditional, yet effective approach allows the model to capture lexical patterns and stylistic markers that are characteristic of each generation method.

Results can be seen in Tables 5 and 6.

Model Architecture	Acc	Rec	F1
Single-Step SVM	0.633	0.659	0.657
Two-Step Pipeline	0.712	0.692	0.695

Table 5
Overall multi-class performance comparison between the single-step SVM baseline and the proposed two-step pipeline (macro averages).

Type of Text	Single-Step Recall	Two-Step Recall
human-written	0.723	0.838
machine-continued	0.637	0.685
machine-polished	0.723	0.758
machine-humanised	0.734	0.707
deeply-mixed	0.470	0.474
Macro recall	**0.659**	**0.692**

Table 6
Per-class recall comparison of the Single-Step SVM and the Two-Step Pipeline.

By combining the binary and multiclass classification stages, our pipeline provides comprehensive classification, first isolating purely human work and then providing a detailed breakdown of how a machine may have been involved in the creation of a given text. The two-step approach yields a significant improvement across all primary metrics over the single-step baseline for all classes except machine-humanised texts.

4.4 Class-Specific Difficulty and Error Analysis

While the macro-averaged metrics demonstrate the superiority of the two-step pipeline, examining the recall on a per-class basis reveals the details around the challenges of AI text detection. As demonstrated in Table 6, the **deeply-mixed** class proved to be the most resilient to detection, yielding a recall of under 0.50 even in the optimized pipeline. This indicates that a mixed text is statistically more likely to be misclassified as one of the other four categories than to be correctly identified. The seamless weaving of human and machine sentences

disrupts the continuous stylistic patterns that both stylometric features and n-gram models rely upon.

Conversely, the **machine-humanized** class-where an LLM text is recursively fed back into an LLM to sound more human-was consistently the easiest to identify across all classical classifiers. This suggests that repeated interactions and iterative prompting do not successfully mask the machine's origin; rather, they compound the artificial stylistic markers, making the resulting text highly distinguishable from authentic human writing.

Limitations

While our proposed dataset and two-step pipeline demonstrate strong improvements in detecting machine-involved text, we acknowledge several limitations in our study. First, the dataset generation process heavily relies on specific proprietary and open-weights models (such as Gemini and OpenAI models). As these LLMs are continuously updated or deprecated, exact reproduction of the generated text distributions may become challenging. Second, our study is strictly limited to the Bulgarian language. While the methodology is intended to be language-agnostic, the specific performance of models like the adapted Binoculars zero-shot detector and the underlying text extraction tools may vary significantly if applied to other languages with different properties. Finally, the stylometric features utilized in the first-stage binary ensemble are sensitive to text characteristics, such as length, part of speech and punctuation frequency, which leaves part of the ensemble more vulnerable to style change attacks.

5. Conclusion

As Large Language Models become increasingly adept at generating and refining Bulgarian text, standard binary detection techniques are no longer sufficient to capture various cases of human-machine collaboration. In this paper, we introduced a novel, multi-category dataset designed to represent varying degrees of machine involvement, ranging from purely human-authored texts to deeply interwoven human-machine content.

Our preliminary evaluations demonstrated the limitations of utilizing a single-step classifier to navigate this complex landscape, primarily due to its tendency to misclassify authentic human writing as machine-assisted. To resolve this, we implemented a hierarchical two-step classification pipeline. By leveraging a binary ensemble combining stylometric features, an adapted Binoculars model, and XLM-RoBERTa, we were able to accurately filtered out purely human-written text in the first stage. This allowed our specialised second-stage multi-class SVM to focus exclusively on categorizing the specific nature of machine generation.

The results confirm that the two-step methodology effectively ensures high recall of human-written content while providing a measurable improvement in overall multi-class accuracy. Future work will focus on expanding the dataset across more diverse text domains and exploring the direct integration of open-source, Bulgarian-specific LLMs into the second stage of the pipeline to further elevate multi-class detection capabilities.

Acknowledgments

This work is partially supported by the project UNITe BG16RFPR002-1.014-0004 funded by PRIDST and EU NextGenerationEU project, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

References

Abassy, Mervat, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, Alham Fikri Aji, Artem Shelmanov, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. LLM-detectAIve: a tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 231–243, Association for Computational Linguistics.

Berbatova, Melania and Filip Ivanov. 2023. An improved Bulgarian natural language processing pipeline. In *Annual of Sofia University "St. Kliment Ohridski", Faculty of Mathematics and Informatics*, volume 10, pages 37–50.

Berbatova, Melania and Yoan Salambashev. 2023. Evaluating hallucinations in large language models for Bulgarian language. In *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, pages 55–63.

Bevendorff, Janek, Daryna Dementieva, Maik Froebe, Bela Gipp, Andre Greiner-Petter, Jussi Karlgren, Maximilian Mayerl, Preslav Nakov, Alexander Panchenko, Martin Potthast, Artem Shelmanov, Efstathios Stamatatos, Benno Stein, Yuxia Wang, Matti Wiegmann, and Eva Zangerle. 2025. Overview of PAN 2025: Generative AI detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection. In *Proceedings of the 47th European Conference on Information Retrieval (ECIR)*.

Bevendorff, Janek, Matti Wiegmann, Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Aarne Talman, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2024. Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*, pages 2486–2506, CEUR-WS.org.

Crothers, Evan N., Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

Csaki, Zoltan, Bo Li, Jonathan Lingjie Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. SambaLingo: Teaching large language models new languages. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 1–21, Association for Computational Linguistics.

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

Jawahar, Ganesh, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain*, pages 2296–2309, International Committee on Computational Linguistics.

Joachims, Thorsten. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98, page 137–142, Springer-Verlag, Berlin, Heidelberg.

Kestemont, Mike, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the cross-domain authorship attribution task at PAN 2019. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2019), Lugano, Switzerland*, volume 2380, pages 1–15.

Kushnareva, Laida, Tatiana Gaintseva, German Magai, Serguei Barannikov, Dmitry Abulkhanov, Kristian Kuznetsov, Eduard Tulchinskii, Irina Piontkovskaya, and Sergey Nikolenko. 2024. AI-generated text boundary detection with RoFT. ArXiv: 2311.08349.

Li, Bohan, Haoliang Qi, and Kai Yan. 2025. Team Bohan Li at PAN: DeBERTa-v3 with R-drop regularization for human-AI collaborative text classification. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain*, volume 4038 of *CEUR Workshop Proceedings*, pages 3763–3769, CEUR-WS.org.

Li, Yafu, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised training shapes translationese in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894.

Macko, Dominik. 2025. mdok of KInIT: Robustly fine-tuned LLM for binary and multiclass AI-generated text detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain*, volume 4038 of *CEUR Workshop Proceedings*, pages 3819–3826, CEUR-WS.org.

Maslo, Andrii and Silvia Gargova. 2025. BuST: A Siamese transformer model for AI text detection in Bulgarian. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 45–52.

Potthast, Martin, Francisco Manuel Rangel-Pardo, Michael Tschuggnall, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. 2017. Overview of PAN'17: Author identification, author profiling, and author obfuscation. In *Lecture Notes in Computer Science*, volume 10456, pages 275–290, Springer.

Rosenthal, Sara, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, editors. 2025. *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Teja, Lekkala Sai, Annepaka Yadagiri, and Partha Pakray. 2025. Team CNLP-NITS-PP at PAN: advancing generative AI detection: Mixture of experts with transformer models. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain*, volume 4038 of *CEUR Workshop Proceedings*, pages 3904–3915, CEUR-WS.org.

Temnikova, Irina, 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, Varna, Bulgaria*, pages 1151–1161.

Tufts, Brian, Xuandong Zhao, and Lei Li. 2025. A practical examination of AI-generated text detectors for large language models. In *Findings of the Association for Computational Linguistics (NAACL 2025), Albuquerque, New Mexico*, pages 4839–4856, Association for Computational Linguistics.

Uchendu, Adaku, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Association for Computational Linguistics.

Voznyuk, Anastasia, German Gritsai, and Andrey Grabovoy. 2025. Team advacheck at PAN: multitasking does all the magic. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038 of *CEUR Workshop Proceedings*, pages 4007–4014, CEUR-WS.org.

Wang, Yuxia, 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 multi-lingual 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, Association for Computational Linguistics.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32, pages 9054–9065, Curran Associates, Inc.

Zeng, Zijie, Lele Sha, Yuheng Li, Kaixun Yang, Dragan Gašević, and Guangliang Chen. 2024. Towards automatic boundary detection for human-AI collaborative hybrid essay in education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22502–22510.

Zheng, Miaoji, Yong Zhong, Fen Liu, Tufeng Xian, Meifang Xie, Weidong Wu, Zhiliang Zhang, and Qiyuan Sun. 2025. StarBERT: a hybrid neural network model for human-AI collaborative text classification. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain*, volume 4038 of *CEUR Workshop Proceedings*, pages 4080–4085, CEUR-WS.org.

Appendices

A. LLM parameters used during generation

Generator Model	Configuration / Hyperparameters
BgGPT 2	`model` = insait/insait-gemma-2-27b-bg-mixed-19847, `temperature` = 0.7, `max_tokens` = 2048, `top_k` = 20, `repetition_penalty` = 1.1, `stop` = [”<end_of_turn>”, ”<eos>”]
Gemini (2.0 Flash & 2.5 Flash)	Default API settings (no explicit parameters passed)
OpenAI o4-mini	Default API settings (no explicit parameters passed)

Table 7
Configuration parameters for the Large Language Models used in the generation of the machine-involved text classes.

B. Prompts and steps for generating texts

Human-written paragraph	Split paragraph into two
“Друга специализирана система...”	“Друга специализирана система... е FACS... Тя позволява кодиране на микромимики...”
“Another specialized system...”	*“Another specialized system... is FACS... It enables the coding of micro-expressions...”*

Table 8
An illustration of how a human-written paragraph is split into two parts.

Title of article	Prompt
"Херменевтика и чуждоезиково обучение"	"Напиши научна статия от около 14 параграфа със заглавие "Херменевтика и чуждоезиково обучение". Не включвай заглавието на статията или заглавия на секции."
"Hermeneutics and Foreign Language Teaching"	*"Write a scientific article of approximately 14 paragraphs with the title "Hermeneutics and Foreign Language Teaching". Do not include the article title or section headings."*
"Стилистика на езика, стилистика на речта или семантическа стилистика"	"Напиши научна статия от около 2564 думи със заглавие "Стилистика на езика, стилистика на речта или семантическа стилистика". Не включвай заглавието на статията или заглавия на секции."
"Stylistics of Language, Stylistics of Speech or Semantic Stylistics"	*"Write a scientific article of approximately 2564 words with the title "Stylistics of Language, Stylistics of Speech or Semantic Stylistics". Do not include the article title or section headings."*

Table 9
A selection of NBU human-written article titles alongside the corresponding prompts given to the large language model o4-mini for generating paragraphs in the "machine-generated" class. The original titles and prompts in Bulgarian as well as English translations are provided.

Instruction	Result
Машинен текст, използван за вход *Machine-generated text used as input*	"Препоръчва се комбиниран подход, който съчетава традиционните техники..." *"A combined approach integrating traditional manual pattern-making..."*
"Промени този параграф да звучи сякаш е писан от човек..." *"Rewrite this paragraph to sound as if it were written by a human..."*	"Комбиниран подход, който обединява класическото ръчно моделиране..." *"A combined approach uniting classical manual pattern-making..."*
"Промени го да звучи сякаш е писан от студент." *"Rewrite it to sound as if it were written by a student."*	"Смятам, че най-добре работи комбинацията..." *"I think the best approach is the combination..."*
"Промени го да звучи по-просто." *"Rewrite it to sound simpler."*	"Най-добре е да комбинирам ръчно моделиране..." *"The best approach is to combine manual pattern-making..."*

Table 10
The sequence of steps applied to a machine-generated text for the "machine-generated, then machine-humanized" class. The original instructions and texts are in Bulgarian; English translations are provided for the reader.

Step	Text
Текст, написан от човек *Human-written text*	"Друга специализирана система... е FACS... FACS представлява..." *"Another specialized system... is FACS... FACS is an anatomically based system..."*
Замаскиран текст *Masked text*	"⟨⟨ПРОПУСНАТА ЧАСТ⟩⟩. FACS представлява..." *"⟨⟨OMITTED PART⟩⟩. FACS is an anatomically based system..."*
Резултат от големия езиков модел *Large language model output*	"Табутата при одита на правното звено водят до игнориране... FACS представлява..." *"Taboos in the audit of the legal unit lead to the ignoring... FACS is an anatomically based system..."*

Table 11
An example of a text with an omitted beginning that the large language model must fill in to produce a "deeply mixed" text.

C. Hyperparameters and training details

Model	Key Hyperparameters & Configuration	Hardware & Runtime
XGBoost	`objective` = binary:logistic, `eval_metric` = logloss, `n_estimators` = 1000, `max_depth` = 6, `learning_rate` = 0.023, `subsample` = 0.722, `colsample_bytree` = 0.843, `gamma` = 0.271, `min_child_weight` = 2, `reg_alpha` = 0.966, `reg_lambda` = 2.617, `early_stopping_rounds` = 50	Intel Core i7 CPU 15 seconds
XLM-RoBERTa	`learning_rate` = 2e-5, `num_train_epochs` = 2, `per_device_train_batch_size` = 16, `per_device_eval_batch_size` = 64, `gradient_accumulation_steps` = 4, `weight_decay` = 0.1, `warmup_steps` = 500, `max_length` = 128, `eval_steps` = 250, `metric_for_best_model` = loss	NVIDIA T4 GPU 2:17 hours
Binoculars	`observer` = SambaLingo-Bulgarian-Base, `performer` = SambaLingo-Bulgarian-Chat	NVIDIA T4 GPU 3:07 hours
Logistic Regression	`max_iter` = 1000	Intel Core i7 CPU 10 seconds
Multi-class SVM	`C` = 0.8, `kernel` = rbf, `gamma` = scale, `class_weight` = balanced, `decision_function_shape` = ovr, `tol` = 0.001, `random_state` = 42	Intel Core i7 CPU 6:45 hours

Table 12
Summary of key hyperparameter configurations and execution environments for the models evaluated in this study.