Team ydong at PAN 2026: Detecting Machine-Generated Text via Multi-Algorithm Compression Distances and Shallow Statistics

Notebook for the PAN Lab at CLEF 2026

Jiacai Yang[1], Qian Dong[1,*]

[1]Foshan University, Foshan, China

Abstract

The rapid proliferation of Large Language Models (LLMs) has blurred the boundaries between human-written and machine-generated text, raising critical security and ethical concerns. Current representative detection methods rely on the perplexity or cross-entropy features of LLMs, which suffer from high inference overhead and poor cross-domain generalization. In this paper, we propose a semantic-agnostic and lightweight detection framework that shifts the detection paradigm from deep linguistic analysis to information-theoretic evaluation. Specifically, we extract a highly compact feature space consisting of Normalized Compression Distance (NCD) based on multiple compression algorithms, including PPMd, zlib, bz2, and lzma. These features capture the inherent structural homogeneity and local redundancy of AI-generated text. We then fuse these features with shallow statistical features, such as N-gram repetition rates and syntactic distribution, to form an 18-dimensional feature vector. With an XGBoost classifier, our framework achieves highly competitive detection performance on benchmark datasets with extremely low computational overhead.

Keywords

Normalized Compression Distance (NCD), AI-generated text, XGBoost

1. Introduction

The rapid advancement of Large Language Models (LLMs) has revolutionized natural language generation, producing text increasingly indistinguishable from human writing. While beneficial for many applications, this capability has also fueled malicious uses—including academic plagiarism, automated misinformation, and spam generation. Consequently, detecting machine-generated text has become a critical priority in the NLP community.

Current state-of-the-art detection methods largely rely on the internal probability distributions of LLMs, typically using perplexity, cross-entropy, or log-rank features. While these zero-shot, semantic-based approaches (e.g., Binoculars) achieve high accuracy, they have two critical limitations. First, they require running massive LLMs for inference, incurring prohibitive computational costs that preclude high-throughput or edge-device deployment. Second, they are inherently biased toward the training distribution of their reference models, causing severe performance drops on text generated by unseen or newer LLMs.

To overcome these challenges, we shift the paradigm from deep linguistic analysis to a semantic-agnostic, information-theoretic evaluation. Unlike human-written text, which exhibits rich semantic variation, machine-generated text—driven by autoregressive next-token prediction—inherently possesses higher predictability, structural homogeneity, and local redundancy. Based on this insight, we introduce a lightweight detection framework that treats text merely as a sequence of discrete symbols. Specifically, we compute Normalized Compression Distance (NCD) using an ensemble of diverse algorithms (PPMd, zlib, bz2, and lzma). Complementing these compression-based features with

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

2429638614@qq.com (J. Yang); dongqian@fosu.edu.cn (Q. Dong)

https://github.com/CREATE1234567890/CompreSense.git (J. Yang);

https://github.com/CREATE1234567890/CompreSense.git (Q. Dong)

shallow statistical features (e.g., N-gram repetition rates and syntactic distribution), we construct a highly compact 18-dimensional feature space. Finally, an XGBoost classifier is trained to make the final decision.

The main contributions of this paper are threefold:

- We propose a novel, semantic-agnostic feature extraction framework that integrates multiple compression algorithms (context-based, dictionary-based, and block-sorting) to capture the compression footprint of AI-generated text.
- We demonstrate that combining these information-theoretic metrics with shallow N-gram statistics forms a highly discriminative 18-dimensional feature vector, mitigating the overfitting issues common in deep learning approaches.
- Extensive experiments on benchmark datasets show that our XGBoost model achieves highly competitive AUC and F1 scores, while operating at a fraction of the inference time and memory required by LLM-based baselines.

2. Related Work

The rapid evolution of LLMs has motivated a family of zero-shot detectors that exploit the internal probability distributions of text. DetectGPT [1] posits that machine-generated text is more sensitive to local perturbations—its log-probability drops more sharply than that of human-written text. Binoculars [2] employs a dual-model architecture, computing cross-perplexity and log-rank ratios between an "observer" and a "performer" LLM. While these semantic-based methods achieve strong accuracy, they face two fundamental limitations. First, they require deploying billion-parameter models (e.g., Llama-3, Falcon) on high-end GPUs, incurring prohibitive computational costs that hinder real-world deployment. Second, they overfit to the statistical biases of their reference models, causing significant performance degradation when detecting text generated by newer or proprietary LLMs. Recent evidence also shows that strong benchmark accuracy can mask real-world failure modes when explainability is considered.

Before the advent of LLMs, authorship verification primarily relied on stylometric features and shallow statistics. Classic baselines such as TF-IDF with SVM or Naive Bayes have been widely used for text classification. While highly efficient, these methods struggle against modern LLMs, which can closely emulate human stylistic traits (e.g., vocabulary richness and sentence length distributions), thus rendering basic N-gram frequencies insufficient for reliable AI text detection. Recent empirical comparisons of TF-IDF with Logistic Regression against Bi-LSTM and DistilBERT report that deep models perform better, while also highlighting challenges from adversarial attacks and rapidly evolving AI writing styles [3]. Earlier semantic network analyses show that even 4-gram generated text remains distinguishable from natural text, reflecting semantic phenomena such as polysemy and synonymy that simple N-gram models fail to capture [4]. However, certain higher-order statistical anomalies—notably excessive bigram and trigram repetition rates—remain enduring artifacts of autoregressive decoding, making them valuable auxiliary signals when properly combined with other features.

Compression algorithms offer a practical approximation of Kolmogorov complexity, providing an information-theoretic lens for measuring text predictability. Sculley and Brodley (2006) pioneered the use of compression models as feature vectors [5], and Halvani et al. (2017) showed the effectiveness of compression distances (e.g., using PPMd) for authorship verification [6]. The core idea is to compare the compressed lengths of individual and concatenated text segments. However, prior work typically employs a single compression algorithm (e.g., PPMd alone), which captures only one form of redundancy (context-based). We address this limitation by integrating multiple complementary algorithms—including dictionary-based (zlib, lzma) and block-sorting (bz2) methods—to construct a more robust and comprehensive feature space.

3. Method

3.1. Framework Overview

Unlike conventional LLM-based zero-shot detectors that rely on token-level log-probabilities, our framework is entirely independent of semantic comprehension, treating input text merely as a sequence of discrete tokens. The pipeline comprises two stages: extracting a highly compact 18-dimensional semantic-agnostic feature vector by combining multi-algorithm compression distances with shallow statistical features, and feeding this vector into an XGBoost classifier to establish a robust detection boundary. Figure 1 provides an overview of the end-to-end pipeline.

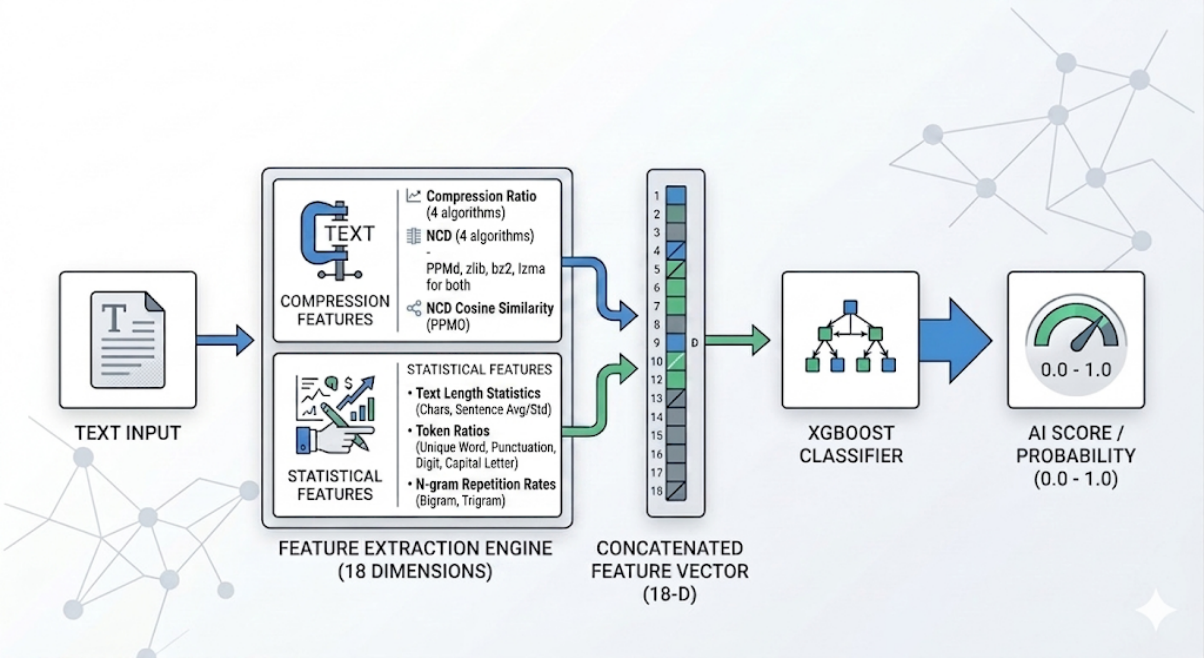

Figure 1: Overview of the MAC-XGBoost pipeline. Text is converted into compression and statistical features, concatenated into an 18-dimensional vector, and classified by XGBoost to output an AI probability.

3.2. Multi-Algorithm Compression Features

Machine-generated texts, driven by autoregressive decoding, inherently exhibit lower information entropy and higher structural predictability than human-written text. To quantify this, we compute Normalized Compression Distance (NCD) and related compression metrics. Given a text T, we split it into two equal halves, T_1 and T_2, to measure its internal self-similarity. Rather than relying on a single compressor, we employ four distinct algorithms that capture complementary redundancy patterns:

- PPMd: A context-based prediction model (Prediction by Partial Matching).
- zlib and lzma: Dictionary-based LZ77 variants that efficiently detect exact string repetitions.
- bz2: A block-sorting algorithm based on the Burrows-Wheeler Transform, sensitive to global symbol permutations.

For each algorithm, let $C(\cdot)$ denote the compressed byte length of a string. The NCD is defined as:

$$\mathrm{NCD}(T_1, T_2) = \frac{C(T_1T_2) - \min(C(T_1), C(T_2))}{\max(C(T_1), C(T_2))}. \tag{1}$$

Here, T_1T_2 denotes the concatenation of T_1 and T_2 (i.e., T). For PPMd, we additionally compute a compression-based similarity metric:

$$S(T_1, T_2) = \frac{C(T_1) + C(T_2) - C(T_1T_2)}{C(T_1) \times C(T_2)}. \tag{2}$$

This approximates normalized mutual information. Finally, the raw compression ratio $C(T)/|T|$ is recorded for each algorithm.

3.3. Shallow Statistics and Repetition Artifacts

Although modern LLMs approximate surface-level human vocabulary, they consistently produce statistical artifacts, most notably in local token repetitions. Let G_n denote the multiset of all n-grams in a document and U_n the set of unique n-grams. We compute repetition rates for bigrams and trigrams as:

$$\mathrm{RR}_n = \frac{|G_n| - |U_n|}{|G_n|}. \tag{3}$$

We further augment the compression features with conventional surface statistics: word count, mean and standard deviation of sentence lengths, unique word ratio, and the proportions of punctuation marks, digits, and uppercase letters. Together, these form the 18-dimensional feature vector.

3.4. XGBoost Classifier

For classification, we adopt an Extreme Gradient Boosting (XGBoost) model with a deliberately shallow architecture. We set the maximum tree depth to 4 (max_depth=4), a learning rate of 0.05, and a total of 300 estimators. This depth constraint acts as a regularizer, mitigating overfitting to dataset-specific artifacts and promoting the learning of fundamental information-theoretic decision boundaries. To handle potential class imbalance, the model automatically adjusts the positive class weight based on the empirical training distribution.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our proposed framework on the benchmark dataset released for the third "Voight-Kampff" Generative AI / LLM Detection task at PAN 2026 [7, 8]. The dataset comprises a balanced mix of human-written texts and machine-generated texts produced by various LLMs under different prompting strategies.

Baselines. We compare our method against three representative baselines provided by the PAN framework:

- Binoculars: A state-of-the-art, zero-shot LLM-based detector utilizing cross-entropy and perplexity scores.
- TF-IDF + SVM: A traditional stylometric baseline using N-gram representations.
- Base PPMd: A single-algorithm compression detector utilizing only context-based compression distances.

Metrics. Detection performance is measured using AUROC, Brier score, C@1, F1-score, and $F_{0.5u}$.

Experimental Environment. Following the PAN shared-task workflow, our runs were developed and evaluated through the TIRA experimentation platform [9]. All experiments, including feature extraction, model training, and inference, were conducted on a standard local machine equipped with an AMD Ryzen 7 5800H CPU with Radeon Graphics (3.20 GHz). Consistent with our framework's lightweight design, the pipeline operates entirely on the CPU without reliance on dedicated GPU VRAM.

4.2. Experimental results and analysis

The results are shown in Table 1. The table shows the performance of our method (MAC-XGBoost) on the test dataset.

Table 1
Score of the model on the test set

Methods	AUROC	Brier	C@1	F1	$F_{0.5u}$
MAC-XGBoost	0.966	0.935	0.912	0.935	0.935
baseline-tf-idf	0.996	0.951	0.984	0.980	0.981
baseline-binoculars-llama-3.1	0.918	0.867	0.843	0.873	0.882
baseline-binoculars-tiny-llama	0.821	0.751	0.627	0.585	0.773
baseline-ppmd	0.786	0.799	0.757	0.812	0.778

Most notably, MAC-XGBoost achieved an AUROC of 0.966 and an F1-score of 0.935. This performance exceeds the state-of-the-art LLM-based zero-shot detector, Binoculars (Llama-3.1), which achieved an AUROC of 0.918 and an F1-score of 0.873. This finding suggests that relying on billion-parameter LLMs for deep linguistic analysis is not necessary. The structural redundancies captured by multi-compression distances provide a discriminative and efficient decision boundary.

Furthermore, compared to the Base PPMd model (AUROC: 0.786, F1: 0.812), which relies solely on a single context-based compression algorithm, our multi-algorithm ensemble improves detection robustness. This supports our hypothesis that combining dictionary-based, context-based, and block-sorting compression algorithms captures a wider spectrum of generative artifacts.

In our experiments, the traditional TF-IDF + SVM baseline achieved the highest raw scores on this dataset (AUROC: 0.996, F1: 0.980), outperforming our proposed MAC-XGBoost. While MAC-XGBoost yields slightly lower absolute scores than this lexical baseline, it offers a fundamentally different perspective for the detection task. Operating within a highly compact 18-dimensional feature space derived from compression distances and shallow statistical patterns, MAC-XGBoost achieves a highly competitive AUROC of 0.966 without any reliance on semantic or lexical content. Moreover, by surpassing Binoculars (AUROC: 0.918), a state-of-the-art LLM detector, our framework demonstrates that information-theoretic features provide a robust and highly efficient alternative foundation for AI text detection.

5. Conclusions

This paper presents MAC-XGBoost, a highly efficient, semantic-agnostic framework for AI authorship verification. Shifting the detection paradigm from deep linguistic analysis to information-theoretic evaluation, we capture the structural redundancies inherent in autoregressive generation. Our method ensembles multiple compression typologies and shallow token statistics into a remarkably compact feature space. Evaluated via an XGBoost classifier, our framework achieves highly competitive classification accuracy—surpassing modern LLM-based zero-shot detectors—while maintaining extreme computational efficiency and mitigating the overfitting risks of traditional stylometric methods. By operating seamlessly on standard CPUs without VRAM constraints, our results confirm that pure compression and statistical algorithms offer a highly scalable and robust solution for modern AI text forensics.

Acknowledgments

We would like to thank Foshan University for supporting this research, and the organizers of the PAN 2026 evaluation campaign for providing the benchmark dataset.

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepSeek in order to: translate text into English. Further, the author(s) used Gemini for Figure 1 in order to: Generate images. After using these

tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, arXiv e-prints (2023) arXiv:2301.11305. doi:10.48550/arXiv.2301.11305. arXiv:2301.11305.

[2] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text, arXiv e-prints (2024) arXiv:2401.12070. doi:10.48550/arXiv.2401.12070. arXiv:2401.12070.

[3] B. Sharimbayev, S. Kadyrov, Text classification for ai generated content with machine learning and deep learning models, in: IEEE 5th International Conference on Smart Information Systems and Technologies (SIST), 2025.

[4] C. Biemann, S. Roos, K. Weihe, Quantifying semantics using complex network analysis, in: Proceedings of COLING 2012, 2012, pp. 263–278.

[5] D. Sculley, C. E. Brodley, Compression and machine learning: A new perspective on feature vectors, Journal of Machine Learning Research 17 (2006) 123–145.

[6] O. Halvani, C. Winter, M. Pohl, On the effectiveness of compression distances for authorship verification, in: Proceedings of the 2017 ACM Workshop on Authorship Investigation, ACM, 2017, pp. 45–52.

[7] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. Potthast, B. Stein, M. N. Ta, Y. Wang, E. Zangerle, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.

[8] J. Bevendorff, R. R. Gunti, J. Karlgren, M. Fröbe, M. Potthast, B. Stein, Overview of the third "Voight-Kampff" Generative AI / LLM Detection Task at PAN and ELOQUENT 2026, in: A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[9] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241.