Team xlbn at PAN 2026: Fusion of Multi-domain Features for Generative AI Authorship Verification

Notebook for the PAN Lab at CLEF 2026

Pinglun Liu[1], Aiguo Wang[1,*]

[1]Foshan University, Foshan 528000, China

Abstract

The boundary between machine-generated and human-written text has become increasingly blurred in the era of large language models. Reliably distinguishing between them is important for mitigating the spread of misinformation and academic misconduct and for preserving content originality and ensuring the credibility of information. To this end, this paper proposes an authorship verification method for the PAN 2026 Voight-Kampff Generative AI Detection task. Specifically, the proposed method combines word-level TF-IDF (term frequency-inverse document frequency), word-boundary character n-gram TF-IDF, and stylometric features to capture complementary textual information. Feature selection then selects a subset of informative features. Finally, a linear classifier with probability calibration is optimized. On the official validation set, the proposed method achieves a mean score of 0.992 and outperforms the official baselines, including linear SVM with TF-IDF, Binoculars, as well as PPMd Compression-based Cosine.

Keywords

Generative AI Authorship Verification, Multi-domain Features, Feature Selection

1. Introduction

Large language models (LLMs) are now able to generate fluent, coherent, and human-like text across a wide range of domains [1]. While it has improved numerous information-processing applications, concerns about academic integrity, misinformation, content moderation, and authorship verification also arise [2, 3, 4]. As generated texts become increasingly natural and stylistically diverse, distinguishing human-written text from AI-generated text has become important yet challenging.

The PAN 2026 Voight-Kampff Generative AI Detection task addresses this problem as a binary detection task [5, 6]. Given an input text, a system is required to determine whether it was written by a human or generated by an AI system, and to assign a confidence score to its prediction. This setting is challenging because detectors may encounter diverse genres, different generation models, and texts whose surface style closely resembles human writing.

Existing approaches to AI generated text detection can be broadly divided into two groups: methods based on neural networks and methods based on machine learning. The former includes supervised neural detectors that fine tune pretrained encoders, such as BERT, DeBERTa, and RoBERTa, for text classification, often with contrastive learning, data augmentation, or ensemble strategies to improve robustness [7, 8, 9]. It also covers statistical detectors that exploit large language models, such as token probabilities, entropy, log rank, and distributional divergence [10, 11, 12]. Although these methods can achieve strong performance, supervised neural detectors usually require labeled data and substantial computational resources, while statistical detectors may depend on access to large language models and often provide black-box decision processes. The latter relies on the feature engineering techniques, where textual features, including TF-IDF, term counts, character patterns, and stylometric statistics, are extracted and then fed into lightweight classifiers [13, 14, 15]. Previous PAN evaluations have shown that such methods remain highly competitive. For example, linear classifiers using TF-IDF and

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

*Corresponding author.

lpinglun@163.com (P. Liu); agwang@fosu.edu.cn (A. Wang)

0009-0000-1404-7633 (P. Liu); 0000-0001-6150-8068 (A. Wang)

term count features can outperform many neural submissions in some cases [13, 14]. This suggests that TF-IDF and stylometric features contain substantial discriminative information, while offering advantages in efficiency, reproducibility, and interpretability. However, existing classical machine learning methods still have several limitations. First, many of them rely on a single view of features, such as lexical TF-IDF alone or a fixed set of stylometric statistics, and they may fail to capture the complementary nature of AI-text signals, which can appear simultaneously in word choice, subword patterns, punctuation usage, sentence structure, and lexical diversity. Second, high-dimensional sparse features may contain redundant and noisy correlations, and linear decision scores are not necessarily reliable confidence estimates.

The above issues motivate us to design an authorship verification method. The proposed method combines word-level TF-IDF, word-boundary character n-gram TF-IDF, and stylometric features to capture complementary information from lexical usage and writing styles. It then performs feature selection to retain stable and discriminative features, and employs a linear classifier with probability calibration to produce confidence scores suitable for the PAN evaluation protocol. This design retains the efficiency and interpretability of feature engineered detectors, while improving their ability to integrate heterogeneous textual signals and support confidence score prediction. The main contributions of this work are as follows:

(1) We propose a detector that fuses word, word-boundary character n-gram, and stylometric features for the PAN 2026 Voight-Kampff Generative AI Detection task. Particularly, feature selection is involved to select a subset of discriminative features, and a linear classifier with calibration score is optimized.

(2) Experimental results show that the proposed model achieves a mean score of 0.992, which outperforms the linear SVM with TF-IDF, Binoculars, and PPMd Compression-based Cosine. Besides, ablation studies concerning the contribution of each component are performed.

2. Related Work

2.1. Neural Network-based Detection

Pretrained Transformer encoders, such as BERT, RoBERTa, and DeBERTa, have been widely used as backbone models for AI-generated text detection. These methods fine tune contextual representations on labeled data for downstream classification [7, 8, 9]. In the PAN setting, this encoder-based paradigm has often been combined with other techniques to improve robustness across genres, domains, and generation sources. The PAN 2024 and 2025 task overviews report studies that fine-tune encoders with data augmentation, regularization and contrastive learning techniques [13, 16]. For example, the RoBERTa detector uses genre embeddings and contrastive learning to capture the semantic differences between human-written and AI-generated texts [17].

Another important line of research uses large language models as probabilistic reference models rather than fine tuning them as classifiers. They assume that machine-generated texts exhibit detectable distributional regularities. For example, GLTR analyze token probabilities, ranks, and entropy to reveal patterns of the generated text [10]. DetectGPT identifies generated passages by measuring the curvature of log probability under perturbations [11], whereas Binoculars compares perplexity behavior between two language models to capture discrepancies between base and instruction-tuned model distributions [12]. LOG-AID extracts surprisal, Jensen-Shannon divergence, entropy, and log-rank features from pretrained language models, and feeds these statistical signals into a logistic regression classifier [18].

Although neural network-based methods show the effectiveness of learned representations for AI-generated text detection, they require task-specific fine-tuning data, access to large pretrained models, and expensive inference. This motivates us to design lightweight alternatives.

2.2. Classical Machine Learning-based Detection

Instead of relying on end-to-end representation learning, Machine Learning models extract lexical, statistical, and stylometric features, and then combine them to build a predictor such as support vector

machines, logistic regression, and gradient-boosted trees. Previous PAN evaluations provide empirical evidence that lexical representations are effective for generative authorship verification. For example, the work at PAN 2024 examined lexical features with logistic regression and SVM, showing that TF-IDF provides discriminative evidence for distinguishing human-written from machine-generated texts [14]. Beyond word-level TF-IDF, several systems have incorporated broader forms of explicit textual evidence. The work in [19] represents texts using word-label correlation values and combines them with statistical measurements. StylOch adopts a modular stylometric pipeline that extracts frequency-based linguistic features from tokenization, dependency relations, part-of-speech tags, named entities, and morphological annotations, uses the gradient-boosted trees for prediction [15]. Compression-based authorship verification methods provide another non-neural perspective by using compression-based similarity to capture regularities related to authorship and text generation [20].

Overall, these studies show that the use of feature engineering remains a competitive direction for AI-generated text detection. However, many existing systems utilize the single-view features such as lexical TF-IDF, word-correlation statistics, stylometric frequencies, and compression-based similarity. In contrast, our study uses word-level TF-IDF to capture lexical preferences, word-boundary character n-gram TF-IDF to represent subword and word-form patterns, and stylometric features to describe writing characteristics.

3. Method

The proposed method follows a three-stage pipeline. It first extracts complementary textual views from the input text, then selects a subset of informative features, and finally trains a calibrated linear classifier to output the confidence score required by the PAN evaluation protocol.

3.1. Multi-domain Feature Extraction

The differences between AI-generated and human-written texts may appear simultaneously in lexical choices, subword patterns, and global writing style. Hence, multi-domain features are first extracted. Specifically, the word-level view use words to capture lexical preferences and local expression patterns, character n-gram TF-IDF view to capture subword and word-form regularities, and the stylometric view describes global writing style through text length, sentence length, lexical diversity, character composition, and punctuation usage.

Given an input text x_i, we encode it into:

$$\Phi(x_i) = [\Phi_{\text{word}}(x_i) \| \Phi_{\text{char}}(x_i) \| \Phi_{\text{style}}(x_i)] \tag{1}$$

where $\|$ denotes feature concatenation. $\Phi_{\text{word}}(x_i)$, $\Phi_{\text{char}}(x_i)$, and $\Phi_{\text{style}}(x_i)$ denote the word-level TF-IDF, character n-gram TF-IDF, and stylometric features, respectively. The character view is a single TF-IDF representation built from character 3- to 5-grams generated within words, with spaces padding word boundaries.

The word-level and character n-gram views adopt the same weighting scheme but differ in their feature units. Let $v \in \{\text{word}, \text{char}\}$ denote the two TF-IDF views, t denote a feature, $c_{i,t}^{(v)}$ denote the count of feature t in x_i, $\text{df}_t^{(v)}$ denote the number of training texts containing feature t, and n is the number of training texts. TF-IDF weights the counts by combining within-text prominence with corpus-level specificity. In the study, the TF-IDF weight is defined as:

$$a_{i,t}^{(v)} = \begin{cases} (1 + \log c_{i,t}^{(v)}) \left(\log \frac{1+n}{1+\text{df}_t^{(v)}} + 1 \right), & c_{i,t}^{(v)} > 0, \\ 0, & c_{i,t}^{(v)} = 0. \end{cases} \tag{2}$$

Each TF-IDF view is then L2-normalized:

$$\Phi_v(x_i) = \frac{a_i^{(v)}}{\|a_i^{(v)}\|_2} \tag{3}$$

Different from TF-IDF features, the stylometric view provides dense document-level statistics that summarize global writing characteristics.

3.2. Feature Selection

Although directly concatenating the multi-domain features provides rich textual information, not all features contribute equally to distinguishing human-written texts from AI-generated ones. To filter out irrelevant and redundant features, feature selection is conducted within the ensemble feature selection framework [21]. The candidate feature set is constructed by concatenating the word-level TF-IDF features, the word-boundary character n-gram TF-IDF features, and the stylometric features.

Feature selection is conducted with logistic regression under two settings: five-fold stratified cross-validation on the full training set and separate training on each valid genre subset. Each fold run retains the top 3,000 features, whereas each genre-specific run retains the top 2,000 features and is applied only to a genre containing at least ten samples in total and examples from both classes. In each run, features are ranked according to the absolute values of the learned coefficients. The selected feature set for setting $r \in R$ is defined as:

$$S_r = \mathrm{TopK}(|\hat{\beta}^{(r)}|) \tag{4}$$

where $\hat{\beta}^{(r)}$ denotes the feature coefficient vector learned by the model in setting r, and $\mathrm{TopK}(\cdot)$ selects the K features with the largest absolute coefficient values.

Each selected feature is assigned one vote, and the final subset retains the features with the highest vote counts. The data after feature selection is noted as:

$$z(x) = \Phi(x)_S \tag{5}$$

3.3. Calibrated Linear Classification

After feature selection, a linear SVM is trained to distinguish human-written texts from AI-generated texts. Given the feature vector $z(x)$, the classifier first produces a margin:

$$m(x) = w^\top z(x) + b \tag{6}$$

where w and b denote the learned weights and bias term, respectively. The margin reflects the relative distance of an instance from the learned decision boundary, but it is not itself a calibrated confidence score. Since the PAN evaluation requires confidence scores, the sigmoid function is used to transform the margin into a probability-like output:

$$f(x) = \sigma(A \times m(x) + B) \tag{7}$$

where A and B are learnable parameters during calibration, and σ denotes the sigmoid function that maps the margin to the interval $[0, 1]$.

4. Experiments

4.1. Experimental Datasets

We evaluate the proposed model on the official Voight-Kampff Generative AI Detection dataset. The task is formulated as a binary classification problem: given an input text, a system outputs a confidence score indicating whether the text was written by a human author or generated by an AI system.

The dataset is provided in newline-delimited JSON format. Each labeled instance in the training and validation splits contains an identifier, the input text, the source model, a binary label, and genre metadata. Two representative records are shown below:

```
{"id": "...", "text": "...", "model": "human", "label": 0, "genre": "essays"}
{"id": "...", "text": "...", "model": "gpt-4o", "label": 1, "genre": "essays"}
```

where label 0 denotes human-written text, whereas label 1 denotes AI-generated text. The genre field describes the textual domain of the instance, and the model field indicates the human or machine source associated with the text. The system was submitted and evaluated through the TIRA platform [22].

4.2. Experimental Setup

In our study, the word-level TF-IDF view uses word unigrams and bigrams, and the character n-gram TF-IDF view uses character 3- to 5-grams constrained within word boundaries. The stylometric view includes 20 numerical features that characterize surface-level writing properties, including text length, lexical diversity, character composition, and punctuation usage. The stylometric features used in this study are summarized in Table 1.

Table 1
Stylometric features used in the proposed method.

View	Features
Stylometric view	number of characters; number of words; number of sentences; average word length; average sentence length measured in words; average sentence length measured in characters; standard deviation of sentence length measured in words; standard deviation of sentence length measured in characters; type–token ratio; ratio of words that occur once; punctuation-character ratio; digit-character ratio; uppercase-character ratio; comma ratio; semicolon ratio; colon ratio; quotation-mark ratio; exclamation-mark ratio; question-mark ratio; newline ratio

The type–token and hapax ratios use word count as the denominator, whereas all other ratios use character count; all stylometric features are standardized before concatenation.

For both TF-IDF views, features occurring in fewer than two training documents are discarded. For the word-level view, terms occurring in more than 95% of the training documents are also removed. Feature selection is done by aggregating feature rankings obtained from five-fold stratified partitions and genre-specific selection runs. Feature selection keeps up to 10,000 features; in the final model, 6,413 features are retained from 1,378,261 candidate features. The retained set comprises 1,838 word-level features, 4,555 character n-gram features, and 20 stylometric features.

The prediction model is a LinearSVC classifier with sigmoid calibration. In the official evaluation setting, systems are expected to produce one confidence score for each input text, with higher scores indicating stronger evidence of AI generation. In our experiments, the official training set is used to learn the detector, while the official validation set is used for reporting experimental results. We do not use any external training data in the experiments.

4.3. Experimental Results

Table 2 summarizes the validation effectiveness of the proposed method and the baselines. We can observe that the proposed method achieves the best results, with the highest arithmetic mean score of 0.992. Compared with the official linear SVM with TF-IDF, the proposed method improves the mean score from 0.978 to 0.992. The ROC-AUC gain is relatively small, increasing from 0.996 to 0.998. By contrast, the Brier score improves markedly from 0.951 to 0.991. The improvements in C@1, F1, and F0.5u further show that the calibrated scores lead to more reliable threshold-based decisions under the PAN evaluation protocol.

Compared with the non-TF-IDF baselines, the proposed method improves the mean score by 0.115 over Binoculars and by 0.206 over PPMd Compression-based Cosine. The consistent gains across ranking, calibration, abstention-aware accuracy, and F-score metrics demonstrate the effectiveness of the joint use of multi-domain features.

Table 2
Experimental results of the proposed method and the baselines. Best results are shown in bold.

Model	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
Linear SVM with TF-IDF	0.996	0.951	0.984	0.980	0.981	0.978
PPMd Compression-based Cosine	0.786	0.799	0.757	0.812	0.778	0.786
Binoculars	0.918	0.867	0.844	0.872	0.882	0.877
ours	**0.998**	**0.991**	**0.989**	**0.991**	**0.990**	**0.992**

4.4. Ablation Study

To evaluate the contribution of each component, ablation experiments are conducted. The results are reported in Table 3.

Table 3
Ablation results of the proposed method. Best results are shown in bold.

Model	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
w/o character n-gram TF-IDF	0.998	0.987	0.985	0.987	0.986	0.989
w/o feature selection	0.998	0.990	0.989	0.990	0.989	0.991
w/o calibration	**0.999**	0.958	**0.994**	0.990	**0.992**	0.987
ours	0.998	**0.991**	0.989	**0.991**	0.990	**0.992**

We can observe that the proposed model achieves the highest mean score of 0.992, confirming the effectiveness of the complete design. Specifically, removing character n-gram TF-IDF decreases the mean score from 0.992 to 0.989, while ROC-AUC remains unchanged. This suggests that character n-gram TF-IDF features provide complementary subword and word-form evidence that mainly improves confidence estimation and threshold-based decisions, rather than substantially changing the model's threshold-independent discriminative ability. Removing the feature selection component leads to a smaller drop, from 0.992 to 0.991, indicating that the use of feature selection helps improve the results by suppressing unstable features. We also observe that without calibration, ROC-AUC, C@1, and F0.5u slightly increase, but Brier drops sharply from 0.991 to 0.958, reducing the mean score to 0.987. This shows that calibration is crucial for reliable confidence scores under the PAN evaluation protocol.

5. Conclusion

This paper presents a multi-domain feature fusion model for generative AI authorship verification. The proposed method combines word-level TF-IDF, word-boundary character n-gram TF-IDF, and stylometric statistics to comprehensively encode the text. Afterwards, to select a subset of informative and high-quality feature, ensemble feature selection is conducted, followed a probability-calibrated linear classifier for prediction. Finally, experimental results on the official validation set show that the proposed method outperforms its competitors, including SVM with TF-IDF, Binoculars, and PPMd Compression-based Cosine.

Declaration on Generative AI

During the preparation of this paper, generative AI tools were used to assist with English language polishing and organization of the manuscript. The authors reviewed and edited all AI-assisted content and take full responsibility for the final publication.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems, 2020.

[2] B. D. Lund, T. Wang, N. R. Mannuru, B. Nie, S. Shimray, Z. Wang, ChatGPT and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing, Journal of the Association for Information Science and Technology 74 (2023) 570–581.

[3] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. E. Tozzi, C. Rizzo, ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health, Frontiers in Public Health 11 (2023) 1166120.

[4] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, D. F. Wong, A survey on LLM-generated text detection: Necessity, methods, and future directions, Computational Linguistics 51 (2025) 275–338.

[5] 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.

[6] 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, Overview of PAN 2026: Voight-Kampff generative AI detection, text watermarking, multi-author writing style analysis, generative plagiarism detection, and reasoning trajectory detection, 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.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[9] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, in: Proceedings of the International Conference on Learning Representations, 2021.

[10] S. Gehrmann, H. Strobelt, A. M. Rush, GLTR: Statistical detection and visualization of generated text, in: Proceedings of the ACL System Demonstrations, 2019.

[11] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, DetectGPT: Zero-shot machine-generated text detection using probability curvature, in: Proceedings of the International Conference on Machine Learning, 2023.

[12] 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, in: Proceedings of the International Conference on Machine Learning, 2024.

[13] J. Bevendorff, M. Wiegmann, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, E. Stamatatos, M. Potthast, B. Stein, Overview of the "Voight-Kampff" generative AI authorship verification task at PAN and ELOQUENT 2024, in: Working Notes of CLEF 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2486–2506.

[14] L. Lorenz, F. Z. Aygüler, F. Schlatt, N. Mirzakhmedova, BaselineAvengers at PAN 2024: Often-forgotten baselines for LLM-generated text detection, in: CLEF Working Notes, 2024, pp. 2761–2768.

[15] J. K. Ochab, M. Matias, T. Boba, T. Walkowiak, StylOch at PAN: Gradient-boosted trees with frequency-based stylometric features, arXiv preprint arXiv:2507.12064 (2025).

[16] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, J. Su, Z. Xie, M. T. Abassy, J. Mansurov,

R. Xing, M. N. Ta, K. A. Elozeiri, T. Gu, R. V. Tomar, J. Geng, E. Artemova, A. Shelmanov, N. Habash, E. Stamatatos, I. Gurevych, P. Nakov, M. Potthast, B. Stein, Overview of the "Voight-Kampff" generative AI authorship verification task at PAN and ELOQUENT 2025, in: Working Notes of CLEF 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[17] J. Yang, K. Yan, Genre-aware contrastive learning for AI text detection: a RoBERTa-based approach, in: Working Notes of CLEF, CEUR Workshop Proceedings, 2025.

[18] S. Titze, O. Halvani, LOG-AID: logit-based statistical features for AI text detection, in: Working Notes of CLEF, CEUR Workshop Proceedings, 2025.

[19] M. Seeliger, P. Styll, M. Staudinger, A. Hanbury, Human or not? light-weight and interpretable detection of AI-generated text, in: Working Notes of CLEF, CEUR Workshop Proceedings, 2025.

[20] O. Halvani, C. Winter, L. Graner, On the usefulness of compression models for authorship verification, in: Proceedings of the 12th International Conference on Availability, Reliability and Security, ACM, 2017, pp. 1–10.

[21] A. Wang, H. Liu, J. Yang, G. Chen, Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data, Computers in Biology and Medicine 142 (2022) 105208.

[22] 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.