Team zhengqiaozeng at PAN 2026: Generative AI Detection Using DeBERTa-v3-large with R-Drop Regularization and Supervised Contrastive Learning

Notebook for the PAN Lab at CLEF 2026

Zhengqiao Zeng, Zhongyuan Han* and Zhankeng Liang

Foshan University, Foshan, China

Abstract

As large language model technologies develop rapidly, these models assist humans in solving complex problems but also pose the challenge of determining whether a given text is machine-generated or human-written. To address this issue, the Voight-Kampff Generative AI Detection 2026 task aims to determine whether a text is AI-generated or human-written. PAN 2026 provides a dedicated dataset for this detection task, where for each given text, the model is required to classify it as machine-generated or human-written. In this study, the pre-trained model DeBERTa-v3-large is employed as the backbone, and a custom loss function that combines cross-entropy loss, R-Drop regularization loss, and supervised contrastive learning loss via a weighted sum is designed for the Voight-Kampff Generative AI Detection task. The model achieves mean scores of 0.952, 0.556, and 0.622 on the pan25-generative-ai-detection, pan26-generative-ai-detection, and eloquent datasets, respectively.

Keywords

DeBERTa-v3-large, PAN 2026, R-Drop, Supervised Contrastive Learning

1. Introduction

With the rapid advancement of artificial intelligence technologies, the capabilities of large language models have been continuously improving. While these models assist humans in solving complex problems, they also bring a new challenge: distinguishing whether a text is human-written or AI-generated is becoming increasingly difficult[1]. The Voight-Kampff Generative AI Detection 2026 task at PAN 2026 is designed to address this issue. This binary classification task requires participants to determine whether a given text is machine-generated or human-written. Compared with test sets from previous years, the new test set also introduces obfuscated content of unknown types to evaluate the robustness of detection systems[1].

In this task, this study builds upon the pre-trained model DeBERTa-v3-large. During training, two forward passes are performed on the same batch to obtain two sets of classification logits and the first token representation from the last hidden state, thereby combining cross-entropy loss, R-Drop regularization loss, and supervised contrastive learning loss for the Voight-Kampff Generative AI Detection task. The experimental results are as follows: the model achieves mean scores of 0.952, 0.556, and 0.622 on the pan25-generative-ai-detection, pan26-generative-ai-detection, and eloquent datasets, respectively. The model performs well on the pan25-generative-ai-detection dataset, while there is still room for improvement on the pan26-generative-ai-detection and eloquent datasets.

2. Related Work

Currently, there are many methods for AI-generated text detection, mainly including zero-shot detection, supervised binary classification, multi-level contrastive learning, and others.

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

** Corresponding author.

zhengqiaozeng@163.com (Z. Zeng); hanzhongyuan@gmail.com (Z. Han); oooo1390329007@163.com (Z. Liang)

0009-0000-5415-349X (Z. Zeng); 0000-0001-8960-9872 (Z. Han); 0009-0009-6124-5494 (Z. Liang)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Eric Mitchell et al. proposed DetectGPT. Their study demonstrates that texts sampled from large language models tend to lie in regions of negative curvature of the log-likelihood function of the model, and based on this observation, they define a curvature-based detection criterion to determine whether a given text is generated by a specific large language model. Specifically, AI-generated texts tend to reside in regions of "negative curvature" of the log-likelihood function of their source model and are extremely sensitive to minor rewrites, whereas human-written texts are less sensitive to minor rewrites[2].

Xun Guo et al. re-examined the task of AI-generated text detection, arguing that the key to this task is not simply classifying a text as human-written or AI-generated, but rather distinguishing the writing styles of different authors. Based on this idea, they proposed DeTeCtive, a detection framework that combines multi-task auxiliary learning and multi-level contrastive learning, aiming to enhance the model's ability to learn representations of different writing styles. In simple terms, each LLM is regarded as an author with a unique writing style, transforming the AI-generated text detection task into a style differentiation task[3].

Although various AI-generated text detection methods have been proposed, Yafu Li et al. constructed a comprehensive testbed by collecting diverse human-written texts and texts generated by different large language models. Experimental results show that accurately distinguishing machine-generated texts from human-written texts remains challenging across various detection scenarios, especially in out-of-distribution settings. A major reason for this challenge is that, as the generative capabilities of large language models improve, the linguistic differences between the two are continually narrowing, and as domains and models become increasingly diverse, detection becomes more and more difficult[4].

Furthermore, Liam Dugan et al. proposed RAID, a large-scale and highly challenging benchmark dataset for machine-generated text detection. Systematic evaluation results based on RAID show that existing detectors are vulnerable to adversarial attacks, variations in sampling strategies, repetition penalty mechanisms, and unseen generation models, leading to a decline in detection performance. This study further indicates that some simple changes, such as modifying sampling strategies or making adversarial modifications to texts, can significantly degrade model performance. In the task of AI-generated text detection, improving model robustness under complex and varying conditions is crucial[5].

3. Method

Voight-Kampff Generative AI Detection 2026 task is a binary classification task. For this task, the pre-trained model used in this study is DeBERTa-v3-large, which consists of 24 Transformer encoder layers with a hidden dimension of 1024. The model was pre-trained on approximately 160 GB of text data [6].

At the beginning of the study, data preprocessing and tokenization were performed, and the samples were then fed into the DeBERTa-v3-large model for training. The core improvement of this study lies in the loss computation stage, where cross-entropy loss, R-Drop regularization loss, and supervised contrastive learning loss are summed in a weighted manner to form the training loss, which is then used to update parameters during backpropagation.

4. Experiment

4.1. Dataset

The model in this study was trained and validated on the training data provided by the PAN 2026 evaluation organizers. [1] While the 2026 task adopts the 2025 training set, different test data are used during the evaluation phase to assess the model's robustness on unseen data. The data are stored in JSONL format, with each record containing five fields: id, text, model, label, and genre. The label field is

[1] https://zenodo.org/records/14962653

set to 0 for human-authored texts and 1 for machine-authored texts. The training set comprises 23,707 texts, and the validation set comprises 3,589 texts.

4.2. Data Processing

For this task, the training set includes 9,101 label 0 samples and 14,606 label 1 samples; the validation set comprises 3,589 samples, of which 2,312 are label 0 and 1,277 are label 1. During training, only the text and label fields were used from the dataset, and other fields were discarded. Data cleaning involved first filtering for correctly formatted samples, followed by preprocessing the text field, mainly removing null characters.

4.3. Model Training

This study fine-tunes the DeBERTa-v3-large model. After data preprocessing and tokenization, the samples are fed into the model. The model performs forward propagation to obtain the logits from the classification head and the hidden states from each layer. The logits are used to compute the cross-entropy loss and the R-Drop regularization loss.

For the cross-entropy loss, the model performs two forward passes with dropout on the same batch of samples, computes the cross-entropy loss between the outputs of each pass and the ground-truth labels, and then takes the average of the two losses as the final cross-entropy loss. In the R-Drop regularization method, the logits obtained from the two forward passes are transformed into probability distributions via the softmax function. The bidirectional KL divergence between the two distributions is then calculated to enforce consistency in the model's outputs. The R-Drop loss weight is set to 1.0.

In the supervised contrastive learning method, this study extracts the first token representation from the last-layer hidden states as the semantic vector of the text. The first token representations from the two forward passes are jointly used to compute the contrastive loss, which encourages representations of samples from the same class to move closer and those from different classes to move apart. The supervised contrastive learning loss weight is set to 0.03, and the temperature coefficient for contrastive learning is set to 0.07.

The final loss function is a weighted sum of the cross-entropy loss, the R-Drop regularization loss, and the supervised contrastive learning loss, and the model parameters are updated through backpropagation. In terms of hyperparameter settings, the number of training epochs was set to 2, the learning rate to 2×10^{-5}, and the effective batch size to 16.

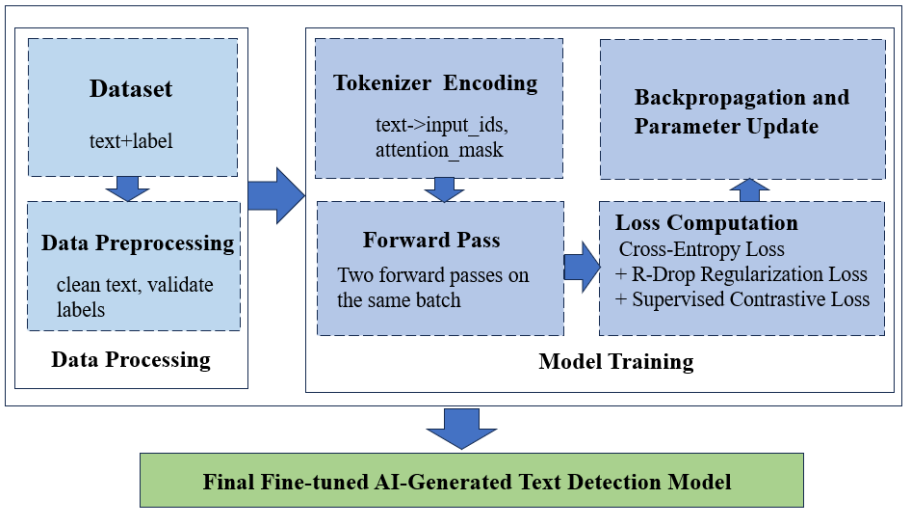

Figure 1: Model Training Process

4.4. Ablation Study

This study designed four groups of ablation experiments and evaluated their performance on the validation set. The four experimental groups adopted the following configurations: cross-entropy(CE) loss only, cross-entropy loss combined with R-Drop regularization loss(CE+R-Drop), cross-entropy loss combined with supervised contrastive learning loss(CE+SCL), and cross-entropy loss combined with both R-Drop and supervised contrastive learning losses(CE+R-Drop+SCL). Regarding hyperparameter settings, the number of training epochs was set to 2, the learning rate was 2×10^{-5}, and the effective batch size was 16. In this study, R-Drop loss and supervised contrastive learning loss were introduced on top of the cross-entropy loss with the expectation of further improving performance. However, the ablation experiments demonstrated that the model performance did not exhibit significant improvement, indicating that the two additional auxiliary losses did not achieve the expected effect. The detailed results are presented in Table 1.

Table 1
Ablation Study

Experiment	ROC-AUC	BRIER	C@1	F_1	$F_{0.5u}$	MEAN
CE	0.99990	0.99632	0.99638	0.99719	0.99629	0.99722
CE+R-Drop	0.99990	0.99529	0.99443	0.99569	0.99389	0.99584
CE+SCL	0.99993	0.99621	0.99582	0.99676	0.99586	0.99692
CE+R-Drop+SCL	0.99992	0.99389	0.99331	0.99483	0.99303	0.99500

5. Results and Analysis

5.1. Results

The official evaluation metrics for this task include six measures: ROC-AUC, which measures the model's ability to distinguish between different classes; Brier score, the complement of the Brier score (itself the mean squared loss between predicted probabilities and true labels), reflecting the quality of probabilistic predictions; c@1, a modified accuracy metric that assigns a score based on the average accuracy of the remaining samples for unanswered (score = 0.5) samples; F_1 score, the harmonic mean of precision and recall, comprehensively measuring the trade-off between them; $F_{0.5u}$, a modified $F_{0.5u}$ measure that places more emphasis on precision and treats unanswered samples (score = 0.5) as false negatives; and mean, the arithmetic mean of all the above metrics, representing the model's overall performance across multiple metrics[1].

This evaluation employed multiple datasets. The results are presented in Table 2[7].

Table 2
Evaluation Results

Dataset	ROC-AUC	BRIER	C@1	F_1	$F_{0.5u}$	MEAN
pan25-generative-ai-detection	0.978	0.931	0.929	0.946	0.976	0.952
pan26-generative-ai-detection	0.753	0.464	0.462	0.439	0.661	0.556
eloquent	0.827	0.46	0.447	0.592	0.782	0.622

The official ranking was based on the pan26-generative-ai-detection dataset, as shown in Table 3[8].

Furthermore, to analyze the model's performance on the pan25-generative-ai-detection dataset (which is identical to the PAN 2025 Subtask 1 dataset), this study conducted a post-hoc retrospective comparison of its results with the official ranking of PAN 2025 Subtask 1. The comparison reveals that the model's score would have achieved second place in that task. The detailed evaluation results are presented in Table 4.

Table 3
PAN 2026

#	Dataset	ROC-AUC	BRIER	C@1	F_1	$F_{0.5u}$	MEAN
30	nai-long-zhan-dui	0.690	0.549	0.477	0.464	0.670	0.570
31	latentauthors	0.750	0.486	0.473	0.460	0.678	0.569
32	zhengqiaozeng	0.753	0.464	0.462	0.439	0.661	0.556
33	spj	0.772	0.437	0.436	0.395	0.620	0.532
34	yihao	0.500	0.750	0.000	0.000	0.000	0.250

Table 4
PAN 2025 Subtask 1 Retrospective Comparison

#	TEAM	ROC-AUC	BRIER	C@1	F_1	$F_{0.5u}$	MEAN
1	Macko	0.995	0.984	0.982	0.989	0.993	**0.989**
	zhengqiaozeng	0.978	0.931	0.929	0.946	0.976	**0.952**
2	Valdez-Valenzuela	0.939	0.902	0.897	0.926	0.960	**0.929**
3	Liu	0.962	0.891	0.889	0.923	0.963	**0.928**
4	Seeliger	0.912	0.898	0.896	0.930	0.959	**0.925**
5	Voznyuk	0.899	0.898	0.898	0.929	0.962	**0.924**
6	Yang	0.930	0.893	0.886	0.920	0.960	**0.923**
7	Zaidi	0.931	0.891	0.887	0.924	0.958	**0.922**
8	hello-world	0.963	0.900	0.897	0.904	0.946	**0.922**
	Baseline TF-IDF SVM	0.963	0.900	0.897	0.904	0.946	0.922
	Baseline Binoculars Llama3.1	0.827	0.856	0.818	0.866	0.907	0.863
	Baseline PPMd CBC	0.644	0.802	0.759	0.817	0.839	0.790

5.2. Analysis

Based on a retrospective comparison against the PAN 2025 leaderboard, the participating systems are listed by their mean scores. The first-place team achieved a score of 0.989, leading the second-place team by 0.060 points. From second to eighth place, the scores were 0.929, 0.928, 0.925, 0.924, 0.923, 0.922, and 0.922, respectively. The difference between the highest and lowest scores in this range is only 0.007 points, indicating that the systems in this interval achieved highly comparable performance.

From the experimental results, the proposed model achieves a mean score of 0.952 on the pan25-generative-ai-detection dataset, surpassing several baseline methods provided by the organizers, including Baseline TF-IDF SVM with a score of 0.922, Baseline Binoculars Llama3.1 with a score of 0.863, and Baseline PPMd CBC with a score of 0.790. This indicates that the fine-tuned model exhibits good task adaptability within the PAN 2025 dataset and its corresponding evaluation framework.

However, according to the official evaluation results on the pan26-generative-ai-detection dataset, the proposed system ranks 32nd. Compared with its performance on PAN 2025, the system's performance has declined. This may be because the datasets of PAN 2025 and PAN 2026 differ significantly, and the model lacks sufficient generalization ability and robustness when facing style imitation, unseen generative models, and text obfuscation methods. The performance of the fine-tuned model on the pan26-generative-ai-detection dataset and its evaluation framework still needs improvement.

6. Conclusion

This study focuses on the Voight-Kampff Generative AI Detection 2026 task. The paper first provides a brief introduction to the task background and the base model used in the Introduction section. Subsequently, the Related Work section reviews the major research advances in AI-generated text detection and the challenges they face. In the methodology section, the main optimization strategies of the proposed model are introduced. The Model Training section describes the dataset, data processing

methods, the detailed training procedure, and the results of ablation studies for this task. In the Results and Analysis section, we report the model's scores on the relevant datasets, and finally, in the results analysis, we provide a retrospective comparison of the model's performance on the PAN 2025 dataset against the 2025 leaderboard, informed by the official ranking rules and baseline models, and also analyze its performance on the PAN 2026 dataset.

Specifically, this study builds an AI-generated text detection model based on the DeBERTa-v3-large pre-trained model. The core improvement lies in a custom loss function, which combines cross-entropy loss, R-Drop regularization loss, and supervised contrastive learning loss via weighted summation to form a new loss function that enhances the training constraints. Under the ranking rules of PAN 2025 Subtask 1, the proposed model obtained a score second only to that of the top-ranked system, surpassing Baseline TF-IDF SVM, Baseline Binoculars Llama3.1, and Baseline PPMd CBC, indicating that the model performs well on the pan25-generative-ai-detection dataset.

However, for the pan26-generative-ai-detection dataset, the official rankings show that there is still room for improvement in the model's performance, and its score on the eloquent dataset is also relatively low. According to the official description, this task introduces additional challenges: large language models were instructed to alter their original writing styles and mimic the expressions of specific human authors. In addition, the test set includes many unknown factors, such as generated texts from new models or unknown text obfuscation methods, to evaluate the robustness of classifiers[1]. Therefore, the performance drop suggests that the model still has limited adaptability to style imitation, unknown generation models, and text obfuscation techniques, and its robustness under complex and unknown perturbations needs further improvement. Future work should analyze the reasons for the performance decline and explore methods to enhance the model's robustness.

7. Acknowledgments

This work is supported by the National Social Science Foundation of China (24BYY080).

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT and DeepSeek-V4 to assist with text polishing and translation. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] 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.

[2] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, Proceedings of Machine Learning Research, PMLR, 2023, pp. 24950–24962. URL: https://proceedings.mlr.press/v202/mitchell23a.html.

[3] X. Guo, Y. He, S. Zhang, T. Zhang, W. Feng, H. Huang, C. Ma, Detective: Detecting ai-generated text via multi-level contrastive learning, in: A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver,

BC, Canada, December 10 - 15, 2024, 2024. URL: http://papers.nips.cc/paper_files/paper/2024/hash/a117a3cd54b7affad04618c77c2fb18b-Abstract-Conference.html.

[4] Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, Y. Zhang, MAGE: machine-generated text detection in the wild, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 36–53. URL: https://doi.org/10.18653/v1/2024.acl-long.3. doi:10.18653/V1/2024.ACL-LONG.3.

[5] L. Dugan, A. Hwang, F. Trhlík, A. Zhu, J. M. Ludan, H. Xu, D. Ippolito, C. Callison-Burch, RAID: A shared benchmark for robust evaluation of machine-generated text detectors, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 12463–12492. URL: https://doi.org/10.18653/v1/2024.acl-long.674. doi:10.18653/V1/2024.ACL-LONG.674.

[6] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA.

[7] 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.

[8] J. Bevendorff, R. R. Gunti, J. Karlgren, M. Fr"obe, 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.