# Team Bohan Li at PAN: DeBERTa-v3 with R-Drop Regularization for Human-AI Collaborative Text Classification

Notebook for the PAN Lab at CLEF 2025

Bohan Li[1,†], Haoliang Qi[2,†] and Kai Yan[3,*,†]

*Foshan University, No. 33 Guangyun Road, Shishan, Nanhai District, Foshan 528225, Guangdong, P.R. China*

**Abstract**

This paper presents our approach to Subtask 2: Human-AI Collaborative Text Classification in the PAN 2025 Voight-Kampff Generative AI Detection challenge. The task focuses on determining the extent to which a text co-authored by humans and artificial intelligence reflects human or machine authorship. The objective is to classify the degree of AI assistance in a given document. In this study, we propose a detection framework that integrates R-Drop regularization with the DeBERTa-v3-base pre-trained language model. The task involves assigning each document to one of six levels of human-AI collaboration, ranging from fully human-written to deeply mixed authorship. To address the challenges of class imbalance and limited training data, we apply random undersampling to high-frequency categories and adopt data augmentation strategies—such as synonym substitution and back-translation—for underrepresented classes. Additionally, R-Drop regularization is introduced during the fine-tuning stage to reduce overfitting and enhance the model's generalization ability on unseen texts. Experimental results show that our proposed model significantly outperforms baseline systems lacking R-Drop and data balancing strategies. On the official test set, our system achieved a macro-level recall of 61.72% and ranked second overall, confirming the effectiveness of the resampling and regularization techniques.

**Keywords**

PAN 2025, Human-AI Collaborative Text Classification, Data Augmentation, R-Drop, DeBERTa-v3

## 1. Introduction

In recent years, large-scale pre-trained language models (LLMs) such as Claude, GPT, and LLaMA have undergone rapid iterations. The resulting advancement in AI-generated content (AIGC) has brought machine-generated texts to a level of fluency and semantic coherence that rivals, and in many cases is indistinguishable from, human-written texts. While these developments have revolutionized applications in dialogue systems, machine translation, and content generation, they have simultaneously posed unprecedented challenges to authorship attribution and content authenticity verification.

To foster progress in this domain, the PAN 2025 [1] shared task on Voight-Kampff Generative AI Detection [2] introduces a more fine-grained subtask: categorizing documents co-authored by humans and AI into six levels of collaboration, ranging from fully human-written to deeply mixed authorship. Existing approaches to AIGC detection generally fall into five categories: watermarking-based tracing, zero-shot perplexity-based detection, fine-tuned language models, adversarial training, and large language models employed as detectors [3]. Each of these strategies emphasizes different aspects such as traceability, unsupervised discrimination, or robustness enhancement. However, under the complex and nuanced six-class setting—particularly when minor human edits are involved—many single-strategy models suffer from limited generalization and vulnerability to adversarial examples.

To address these challenges, we propose a classification framework that combines the DeBERTa-v3-base [4, 5] pre-trained model with R-Drop regularization [6]. DeBERTa-v3 enhances long-range

*CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain*

[*] Corresponding author.

[†] These authors contributed equally.

137765525@qq.com (B. Li); haoliang.qi@gmail.com (H. Qi); yankai@fosu.edu.cn (K. Yan)

0009-0000-9242-0130 (B. Li); 0000-0003-1321-5820 (H. Qi); 0000-0002-4960-7108 (K. Yan)

dependency modeling through disentangled attention and improved decoder masking, enabling more accurate representation of subtle stylistic variations [4, 5]. R-Drop introduces dual forward passes during training and minimizes the Kullback–Leibler divergence between outputs, thereby reducing overfitting and encouraging decision boundary smoothness [6]. In addition, we apply a training data resampling strategy that combines undersampling of majority classes with multi-strategy data augmentation [7] for minority classes, including synonym substitution and back-translation, to improve representation diversity.

By integrating the DeBERTa-v3 pre-trained model, R-Drop regularization, and data balancing and augmentation techniques, we construct a six-way classifier to assess the degree of human-AI collaboration in text. Experimental results on the official development set demonstrate that our model significantly outperforms baseline models without R-Drop or data balancing, confirming the effectiveness of the proposed approach.

## 2. Background

With the rapid development and widespread deployment of large language models (LLMs) such as Claude, GPT, and LLaMA, AI-generated content (AIGC) detection has emerged as a critical research direction for ensuring content credibility and copyright compliance. This task is typically formulated as a text classification problem. In this section, we provide a structured overview of current AIGC detection strategies, which can be broadly categorized into watermarking techniques, zero-shot detection models, supervised learning-based detectors, adversarial and robustness-oriented methods, and LLM-as-detector paradigms [3].

**Watermarking Techniques:** Watermarking techniques embed verifiable statistical fingerprints into generated texts during inference, such as controlled token distributions or congruence-based constraints, enabling downstream statistical tests to verify content provenance efficiently [8]. These methods offer fast inference and low false-positive rates, and—if enforced at the generation source—can provide near-deterministic traceability. However, watermarks are often vulnerable to dilution through post-processing steps such as clipping, translation, or rewriting, and are ineffective against unauthorized APIs or unknown-source texts.

**Zero-Shot Detection Models:** Zero-shot approaches do not rely on labeled training data; instead, they distinguish human- and machine-authored texts using statistical signals such as perplexity, entropy, or n-gram rarity. Tools like GLTR [9]and DetectGPT [10], for example, examine anomalies in token confidence distributions to detect machine authorship. These models are naturally domain- and language-agnostic, but their effectiveness diminishes on high-quality or human-refined texts. Moreover, some variants incur substantial computational costs due to repeated perturbations or multiple forward passes.

**Supervised Learning Approaches:** Supervised methods leverage pre-trained language models such as BERT [11], RoBERTa [12], and DeBERTa [4, 5], which are fine-tuned on annotated human-AI hybrid corpora to capture deep semantic and syntactic distinctions. These models generally perform well in single-domain, large-scale, and balanced datasets, and can be extended to multi-level classification tasks. However, their generalization is often limited by the training distribution, leading to overfitting on out-of-domain inputs or evasive edits, and their performance is highly sensitive to underrepresented classes.

**Adversarial Learning Methods:** This line of work enhances model robustness by generating adversarial examples, incorporating consistency regularization (e.g., R-Drop), or employing contrastive loss functions. For example, RADAR [13]and OUTFOX [14]. Empirical studies show that adversarial training can substantially reduce the success rate of paraphrasing or syntactic evasion attacks, while also enabling models to estimate confidence or uncertainty in predictions. Nevertheless, these methods typically require carefully designed adversarial strategies, incur high training costs, and their gains may be limited when synthetic adversarial samples diverge significantly from real-world attacks.

**LLMs as Detectors:** Large language models (LLMs) can assess authorship by utilizing prompts that

**Table 1**
Label distribution statistics before and after balancing.

| Label Category | Train | Original % | Balanced | Balanced % |
|---|---|---|---|---|
| Machine-written, then machine-humanized | 91,232 | 31.6% | 40,000 | 24.24% |
| Human-written, then machine-polished | 95,398 | 33.0% | 40,000 | 24.24% |
| Fully human-written | 75,270 | 26.1% | 40,000 | 24.24% |
| Human-initiated, then machine-continued | 10,740 | 3.7% | 20,000 | 12.12% |
| Deeply-mixed text (human + machine) | 14,910 | 5.2% | 20,000 | 12.12% |
| Machine-written, then human-edited | 1,368 | 0.4% | 5,000 | 3.03% |
| **Total** | 288,918 | 100% | 165,000 | 100% |

frame the detection task. Early results were erratic and highly prompt-sensitive. However, in-context learning (ICL) [15]has improved stability by embedding a few curated input–label examples within the prompt. Experimental findings demonstrate that the ICL strategy outperforms both traditional zero-shot methods and RoBERTa-based detectors.

## 3. System Overview

In this section, we present the experimental model and methodology. Our approach is built upon the DeBERTa-v3-base pre-trained language model, enhanced by the incorporation of R-Drop regularization. In addition, we apply data balancing and augmentation strategies to the training set, which comprises 288,918 samples. These methods are designed to improve the model's generalization ability and enhance its stability during inference.

### 3.1. Data Balancing and Augmentation

Large-scale imbalanced corpora often lead classifiers to overfit to majority classes, resulting in poor performance on underrepresented categories—particularly classes 3, 4, and 5 in the six-way classification task. To ensure the model captures fine-grained patterns of human-AI collaboration, we construct a two-stage preprocessing pipeline consisting of **undersampling** for majority classes and **multi-strategy augmentation** for minority classes.

The three most frequent classes (0–2) together account for 90.7% of the total training data. Direct training on such skewed distributions would severely bias the decision boundaries. Based on preliminary assessments of class difficulty and model capacity, we define a target class distribution of **40k:40k:40k:20k:20k:5k**, which significantly increases the weight of rare categories while avoiding excessive pruning of majority-class instances. As shown in Table 1.

For majority classes (0, 1, and 2), we apply undersampling by fixing the random seed `random.seed(42)` and randomly sampling 40,000 representative and diverse instances from each class.

For minority classes (3, 4, and 5), we apply multi-strategy augmentation. Specifically, classes 3, 4, and 5 are expanded using random oversampling combined with six data augmentation strategies:

- **Random Swap**: Randomly swaps 15% of word positions to increase syntactic diversity.
- **Random Deletion**: Deletes words with a probability of 0.1 to simulate abbreviation and compression.
- **Synonym Replacement**: Replaces selected non-stopwords with their synonyms (retrieved via embedding or lexical databases) to preserve semantics while diversifying expression.
- **Back-Translation**: Introduces structural variation through machine translation and reconstruction.
- **Sentence Shuffle**: Retains the first and last sentences while shuffling intermediate ones to simulate paragraph-level rewriting.

- **EDA Combination**: Applies multiple Easy Data Augmentation (EDA) operations (e.g., swap, synonym replacement) sequentially to generate highly heterogeneous variants.

These methods are applied with uniform random selection. As a result, classes 3 and 4 are each augmented to 20,000 instances, while the most underrepresented class 5 is expanded from 1,368 to 5,000 instances.

This “undersampling + multi-strategy augmentation” pipeline effectively mitigates distributional bias and provides a balanced and diverse input space for subsequent fine-tuning with the DeBERTa model and R-Drop regularization.

## 3.2. R-Drop Regularization

During fine-tuning of the DeBERTa-v3-base model, we adopt R-Drop (Regularized Dropout) to impose a consistency constraint on the conventional dropout mechanism, aiming to mitigate overfitting and enhance the model’s robustness in distinguishing fine-grained labels. Unlike traditional dropout, which performs a single forward pass with stochastic masking, R-Drop conducts two independent forward passes with different dropout masks on the same input batch and minimizes the Kullback–Leibler (KL) divergence between their output distributions. This encourages consistency between the two predictions and explicitly reduces discrepancies among sub-network outputs.significantly lowering the model’s reliance on specific neuron co-activations and improving generalization.

To formalize the R-Drop loss, let the input sample be denoted as $x_i$ with its ground-truth label $y_i$. Under two independent dropout masks, the model produces predictive distributions $P_{\theta_1}(y_i \mid x_i)$ and $P_{\theta_2}(y_i \mid x_i)$. The R-Drop loss combines dual cross-entropy with a symmetric Kullback–Leibler (KL) divergence term:

$$\mathcal{L}_{\text{R-Drop}} = \frac{1}{2}\left[\text{CE}(y_i, P_{\theta_1}) + \text{CE}(y_i, P_{\theta_2})\right] + \alpha \cdot \frac{1}{2}\left[D_{\text{KL}}(P_{\theta_1} \parallel P_{\theta_2}) + D_{\text{KL}}(P_{\theta_2} \parallel P_{\theta_1})\right] \tag{1}$$

## 3.3. Supervised Fine-Tuning

We adopt DeBERTa-v3-base as the backbone model. Owing to its disentangled attention mechanism and enhanced masked decoder, DeBERTa-v3-base demonstrates strong capabilities in modeling complex semantic dependencies and capturing positional relationships, significantly improving contextual understanding and structural representation.

To construct a balanced training dataset, we perform undersampling on majority classes (labels 0, 1, and 2) and apply data augmentation to minority classes (labels 3, 4, and 5), resulting in a well-proportioned dataset of approximately 165,000 instances.

Subsequently, the model is fine-tuned with the R-Drop regularization technique. For each training batch, two independent forward and backward passes are performed under different dropout masks, and the Kullback–Leibler (KL) divergence between the two output distributions is calculated as a regularization term. This encourages predictive consistency and helps reduce uncertainty, thus enhancing model robustness.

To prevent overfitting, we apply an early stopping strategy: training terminates once the validation loss ceases to decrease. The model is evaluated using the official metrics—recall and F1 score—and the best-performing checkpoint across all training epochs is retained. Final performance is assessed on the held-out test set.

As shown in Algorithm 1, the complete fine-tuning procedure of R-Drop applied to DeBERTa-v3-base is detailed.

**Algorithm 1** Fine–Tuning DeBERTa-v3 with R-Drop Regularization

```
Require: Raw training set $\mathcal{D}_{\text{train}}$, raw development set $\mathcal{D}_{\text{dev}}$
Require: Pre-trained model DeBERTa-v3-base with parameters $\theta$
Require: Undersample size $U$=40k for classes 0–2, augmentation targets $T_3$=$T_4$=20k, $T_5$=5k
Require: Hyper-parameters: learning rate $\eta$, batch size $B$, epochs $E$, R-Drop weight $\alpha$=1.0, random seed $s$
Ensure: Fine-tuned model $\theta^\star$ with best validation performance
Set random seed $s$; initialize tokenizer and optimizer with $\eta$                ▷ Stage 1: Data Balancing
Split $\mathcal{D}_{\text{train}}$ by label $\ell \in \{0, \dots, 5\}$: $\{\mathcal{D}_\ell\}$
Undersample majority classes: $\mathcal{D}_\ell \leftarrow \text{Sample}(\mathcal{D}_\ell, U)$ for $\ell \in \{0, 1, 2\}$
Augment minority classes ($\ell \in \{3, 4, 5\}$) with strategies random_swap, random_deletion, back_translation, sentence_shuffle, EDA_combination until $|\mathcal{D}_3|$=$|\mathcal{D}_4|$=$T_3$ and $|\mathcal{D}_5|$=$T_5$
$\mathcal{D}^{\text{bal}}_{\text{train}} \leftarrow \bigcup_{\ell=0}^{5} \mathcal{D}_\ell$; shuffle with seed $s$                ▷ Stage 2: Tokenization
Tokenize $\mathcal{D}^{\text{bal}}_{\text{train}}$ and $\mathcal{D}_{\text{dev}}$ using max length 512                ▷ Stage 3: R-Drop Fine-Tuning
for epoch $e = 1$ to $E$ do
    for each mini-batch $(X_b, Y_b) \subset \mathcal{D}^{\text{bal}}_{\text{train}}$ do
        Forward pass #1: $(\mathbf{z}_1, \text{CE}_1) \leftarrow M_\theta(X_b, Y_b)$
        Forward pass #2: $(\mathbf{z}_2, \text{CE}_2) \leftarrow M_\theta(X_b, Y_b)$                ▷ independent dropout masks
        Compute symmetric KL loss: $\text{KL} = \frac{1}{2}[D_{\text{KL}}(\mathbf{z}_1 \parallel \mathbf{z}_2) + D_{\text{KL}}(\mathbf{z}_2 \parallel \mathbf{z}_1)]$
        Total loss: $\mathcal{L} = \frac{1}{2}(\text{CE}_1 + \text{CE}_2) + \alpha\,\text{KL}$
        Back-propagate $\nabla_\theta \mathcal{L}$; update $\theta$ with AdamW
    end for
    Evaluate on dev subset; save $\theta$ if $\text{F1}_{\text{macro}}$ improves
    if validation loss has not decreased for $p$ consecutive epochs then
        break                ▷ early stopping
    end if
end for                ▷ Stage 4: Final Evaluation
Load best checkpoint $\theta^\star$; evaluate on full dev/test set and report accuracy, recall, $\text{F1}_{\text{macro}}$, $\text{F1}_{\text{micro}}$
return Fine-tuned model parameters $\theta^\star$
```

# 4. Experiments

## 4.1. Experimental settings

**Model** We adopt **DeBERTa-v3-base**[1] as the encoder, given its disentangled self-attention and enhanced mask decoder, which have demonstrated strong performance on long-sequence classification tasks.

**Input preprocessing** All documents are tokenised with the original DeBERTa WordPiece tokenizer. Sentences exceeding 512 tokens are truncated, while shorter ones are padded on-the-fly with the special <pad> token.

**Hyper-parameters** Table 2 lists the full configuration used in every run. The R-Drop weight $\alpha$ is set to **1.0** after a coarse logarithmic search in $\{0.05, 0.2, 1, 5, 10\}$.

**Evaluation protocol** After each epoch, we save a checkpoint and we evaluate the model on the development set to obtain timely feedback. During this evaluation, we compute metrics including macro-F1, micro-F1, accuracy, and macro-recall, which facilitate the selection and preservation of the best-performing model.

[1] 12 transformer layers, hidden size 768, 12 attention heads.

**Table 2**
Hyper-parameter setup for all experiments.

| Parameter | Value | Note |
|---|---|---|
| Max sequence length | 512 | Prevents excessive truncation |
| Learning rate | $2 \times 10^{-5}$ | Linear decay |
| Batch size (train/eval) | 16 / 16 | |
| Epochs | 10 | Early stop patience 2 |
| Weight decay | 0.01 | $L_2$ regularisation |
| R-Drop weight $\alpha$ | 1.0 | Consistency strength |
| Dropout probability | 0.10 | As in the original PLM |
| Optimizer | AdamW | $\beta_1{=}0.9,\ \beta_2{=}0.999$ |
| Scheduler | Linear decay | Step-wise update each batch |
| Precision | FP16 | NVIDIA AMP |

**Table 3**
Performance comparison between our system and the official baseline on the PAN 2025 test set.

| **Team Name** | **Recall (Macro)** | **F1 (Macro)** | **Accuracy** |
|---|---|---|---|
| lbh-1130 | 61.72% | 61.73% | 69.28% |
| Baseline | 48.32% | 47.82% | 57.09% |

### 4.2. Result

As shown in Table 3, our model clearly outperforms the baseline, achieving higher scores on *Recall (Macro)*, *F1 (Macro)*, and *Accuracy*.

## 5. Conclusion

This paper presents our work on Subtask 2: Human-AI Collaborative Text Classification of the PAN 2025 Voight-Kampff Generative AI Detection challenge. We built our system based on the DeBERTa-v3-base model, enhanced with R-Drop regularization and data balancing and augmentation techniques, in order to improve the model's generalization and robustness. Our approach ultimately achieved a strong second-place result in the task. Comparative experiments demonstrated that incorporating R-Drop during training positively contributed to the overall performance of the model. However, due to the lack of more refined data processing and our still-limited understanding of the model architecture, the system did not reach its full potential, leaving room for further improvement. In future work, we plan to focus on deeper architectural enhancements to further improve performance.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62276064)

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References

[1] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[2] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsivgun, J. Su, Z. Xie, M. 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: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[3] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, L. S. Chao, A survey on llm-generated text detection: Necessity, methods, and future directions, 2024. URL: https://arxiv.org/abs/2310.14724. arXiv:2310.14724.

[4] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced BERT with disentangled attention, CoRR abs/2006.03654 (2020). URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.

[5] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, CoRR abs/2111.09543 (2021). URL: https://arxiv.org/abs/2111.09543. arXiv:2111.09543.

[6] X. Liang, L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T. Liu, R-drop: Regularized dropout for neural networks, CoRR abs/2106.14448 (2021). URL: https://arxiv.org/abs/2106.14448. arXiv:2106.14448.

[7] J. W. Wei, K. Zou, EDA: easy data augmentation techniques for boosting performance on text classification tasks, CoRR abs/1901.11196 (2019). URL: http://arxiv.org/abs/1901.11196. arXiv:1901.11196.

[8] C. Gu, C. Huang, X. Zheng, K.-W. Chang, C.-J. Hsieh, Watermarking pre-trained language models with backdooring, 2023. URL: https://arxiv.org/abs/2210.07543. arXiv:2210.07543.

[9] S. Gehrmann, H. Strobelt, A. M. Rush, GLTR: statistical detection and visualization of generated text, CoRR abs/1906.04043 (2019). URL: http://arxiv.org/abs/1906.04043. arXiv:1906.04043.

[10] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, 2023. URL: https://arxiv.org/abs/2301.11305. arXiv:2301.11305.

[11] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[12] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[13] X. Hu, P.-Y. Chen, T.-Y. Ho, Radar: Robust ai-text detection via adversarial learning, 2023. URL: https://arxiv.org/abs/2307.03838. arXiv:2307.03838.

[14] R. Koike, M. Kaneko, N. Okazaki, Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples, 2024. URL: https://arxiv.org/abs/2307.11729. arXiv:2307.11729.

[15] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, B. Chang, X. Sun, L. Li, Z. Sui, A survey on in-context learning, 2024. URL: https://arxiv.org/abs/2301.00234. arXiv:2301.00234.